Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

A SYSTEMATIC REVIEW OF CAN BUS INTRUSION DETECTION SYSTEMS: GAPS IN EXPLAINABILITY, AUTONOMOUS RESPONSE, AND SOC INTEGRATION

Muhammad Hamid[*1], Anfal Tariq[2], Iftikhar Rasheed[3], Asjad Amin[4]

[*1,2,3,4]*Faculty of Engineering and Technology, The Islamia University of Bahawalpur, Pakistan*

hamidcit14@gmail.com, anfal.ch212@gmail.com, iftikhar.rasheed@iub.edu.pk, asjad.amin@iub.edu.pk

DOI: https://doi.org/10.5281/zenodo.20521105

Keywords
CAN bus, intrusion detection systems, automotive cybersecurity, explainable AI, SOC, autonomous response. Controller Area Network (CAN), intrusion detection systems (IDS), automotive cybersecurity, explainable AI (XAI), SHAP, LIME, counterfactual explanations, security operations center (SOC), autonomous response, systematic review.

Article History
Received on 20 March 2026
Accepted on 12 May 2026
Published on 27 May 2026

Corresponding Author: *
Muhammad Hamid*

Abstract

The Controller Area Network (CAN) bus is the primary communication network in an in-vehicle environment today, despite not having authentication, encryption or access control, rendering it highly susceptible to cyber-physical attacks. In recent years, many machine learning and deep learning based intrusion detection systems (IDS) have been proposed for intrusions in CAN traffic, which are often able to provide a high detection accuracy in the benchmark data sets. Most models, however, are detection only, black-box models, are not very explainable, and have not been integrated into Security Operations Center (SOC) processes or have the ability to respond on their own. This review critically examines the scope of the existing CAN-IDS work in terms of detection capability, explainability, automation of responses, diversity of datasets, and evaluation approaches. It reveals critical areas that can obstruct deployment in the field such as lack of generalisation between vehicle platforms, and lack of operationally integrated defence strategies. The paper suggests future research paths for explainable integration, autonomous responding, monitoring with SOC focus, multi-dataset evaluation, and safety-driven deployment. The review outlines a path forward for enhancing CAN ID to become active, explainable and operationally integrated protection.

1. INTRODUCTION

Modern vehicles have evolved from purely mechanical machines into highly complex cyber-physical systems composed of more than seventy distributed electronic control units (ECUs) interconnected through multiple in-vehicle networks [1]-[2]. Introduced by Bosch in 1986 and standardized as ISO 11898, the Controller Area Network (CAN) remains the de facto backbone for powertrain, chassis, and body-control communications due to its deterministic arbitration, low cost, and proven robustness in automotive environments [3].

However, CAN was conceived when vehicles were considered closed systems; consequently, the protocol lacks message authentication, payload encryption, sender identification, and access control [4]-[5].

The rapid expansion of the vehicular attack surface has been driven by telematics, infotainment systems, on-board diagnostics (OBD-II) ports, Bluetooth, Wi-Fi, cellular modems, and over-the-air (OTA) update channels. [6]-[8].Public demonstrations such as the 2015 Jeep Cherokee remote takeover [9] and subsequent attacks on Tesla, BMW, and Toyota platforms [10]-[11] showed that adversaries can exploit peripheral entry points to reach the CAN bus and manipulate steering, braking, and transmission. As a result, automotive cybersecurity has become a regulatory priority under UNECE WP.29 R155/R156 and ISO/SAE 21434 [12].

In response, the research community has produced extensive work on machine learning (ML) and deep learning (DL)–based intrusion detection systems (IDS) for CAN traffic. Tree-based classifiers [13]-[16],recurrent neural networks (RNN/LSTM/GRU) [17]-[20],convolutional neural networks (CNN) [21]-[23], hybrid models [24]-[25], and autoencoders [26] have been applied to attack classes including denial-of-service (DoS), fuzzy injection, spoofing, replay, masquerade, and suspension.Reported detection accuracies routinely surpass 99%, making CAN IDS seem like a solved problem. However, this article attempts to debunk this myth. A systematic review of recent literature shows that there is an overly restrictive paradigm in this field, which is limited by five factors:

Detection is considered the end-stage functionality, without any autonomous preventive or reactive measures.

Decision-making processes are based on black-box approaches, lacking explanation for individual alerts.

The IDS outputs are mere alerts that do not integrate into the SOC process.

The training and evaluation of the model are done using one dataset which is outdated.

Software-based protection mechanisms (e.g., ID whitelisting, rate limiting, frame blocking, ECU bus-off) are rarely incorporated into end-to-end pipelines.

The contributions of this review are threefold:

A structured synthesis of contemporary CAN-IDS research mapped across eleven capability dimensions.

Formal identification of fundamental research gaps spanning detection, explainability, response automation, dataset diversity, and operational integration.

A forward-looking research roadmap toward explainable, autonomous, and SOC-oriented CAN intrusion detection.

The remainder of this paper is organized as follows. Section II reviews CAN architecture and design weaknesses. Section III presents the threat model and attack taxonomy. Section IV critically reviews existing CAN-IDS literature. Section V discusses the black-box nature of ML/DL-based IDS. Section VI surveys explainable AI techniques applicable to IDS. Section VII introduces SOC concepts for automotive security. Section VIII addresses

dataset and evaluation challenges. Section IX presents a comparative capability matrix. Section X consolidates the key research gaps. Section XI outlines future research directions toward an autonomous XAI-SOC framework. Section XII concludes the paper.

LITERATURE REVIEW: CAN BUS ARCHITECTURE AND SECURITY WEAKNESSES

This section reviews the Controller Area Network (CAN) architecture and its security limitations, synthesizing key findings from automotive cybersecurity research. The literature consistently shows that CAN's vulnerabilities are structural and protocol-inherent, motivating the development of intrusion detection systems (IDS) as the primary defense mechanism.

CAN Bus Architecture Overview

The Controller Area Network (CAN) is a multi-master, message-oriented broadcast protocol operating over a differential twisted-pair medium. It supports data rates up to 500 kbps (Classical CAN) and uses 11-bit or 29-bit identifiers that define both message content and arbitration priority, where lower values have higher priority.

The protocol employs nondestructive arbitration, ensuring deterministic real-time communication. A typical frame includes a Data Length Code (0–8 bytes), payload, CRC field, and acknowledgment bits [3].

Early works assume three design conditions: (i) trusted nodes, (ii) isolated networks, and (iii) no adversarial presence [1], [5]. These assumptions no longer hold in modern connected vehicles.

Structural Security Weaknesses in CAN

Extensive literature identifies five fundamental and universal vulnerabilities in CAN systems [2], [4], [33].

Core CAN Security Weaknesses

Weakness	Description	Key Impact
No authentication	Any ECU can spoof identifiers	ECU impersonation
No encryption	All data transmitted in plaintext	Eavesdropping
No access control	Broadcast to all ECUs	Lateral movement
Priority-based DoS	ID 0x000 dominates arbitration	Bus starvation
Error abuse	Forced ECU bus-off state	ECU shutdown

These weaknesses have been experimentally validated across real vehicles and platforms [32], [50], [51].

Evolution of CAN Security Research

Research on CAN security has progressed through four phases:

Research Evolution Timeline

Phase	Period	Focus
Discovery	2004–2010	Vulnerability identification
Exploitation	2010–2015	Vehicle attacks
Remote Attacks	2015–2018	Telematics exploitation
Defense (IDS)	2018–2025	ML/DL-based detection

Analysis of Key Vulnerabilities

Authentication and Confidentiality Issues

CAN lacks sender authentication and encryption, enabling both impersonation and passive monitoring. Studies show that ECUs can be spoofed and vehicle state can be fully reconstructed using low-cost hardware [36], [50].

Access Control Limitations

The broadcast nature of CAN enables cross-ECU communication without restrictions. Gateway ECUs provide partial isolation, but studies show frequent bypass via diagnostic services [48], [65].

Denial-of-Service via Priority Exploitation

The arbitration mechanism allows attackers to flood the bus using high-priority identifiers (e.g., 0x000), causing rapid bus saturation (<10 ms) [48], [73].

Error Handling Abuse

Error confinement mechanisms can be exploited to force ECUs into bus-off state, effectively disabling critical vehicle functions within seconds [35], [74].

Cryptographic Countermeasures and Limitations

Although cryptographic solutions such as AUTOSAR SecOC and CAN authentication protocols exist, their adoption is limited by three key constraints:

Limitations of Cryptographic Solutions

Constraint	Impact
8-byte payload limit	Insufficient space for security data
Real-time latency	Violates control loop timing
Legacy compatibility	Incompatible with existing ECUs

These limitations make full-scale deployment across existing vehicle fleets impractical [50], [86].

IDS Capability Comparison

IDS Capability Summary

Work	DL	MA	XAI	AR	MD
Song et al. 2020 [21]	✓	✓	–	–	–
Hanselmann et al. 2020 [26]	✓	✓	–	–	–
Mehedi et al. 2021 [19]	✓	✓	–	–	–
Verma et al. 2024 [29]	✓	✓	–	–	✓
Alkhatib et al. 2023 [49]	✓	✓	✓	–	–
M. H. Shahriar et al. 2024 [50]	✓	✓	✓	–	–
J. Lee et al. 2025 [56]	✓	✓	✓	–	–

DL = Deep Learning, MA = Multi-Attack, XAI = Explainable AI, AR = Auto Response, MD = Multi-Dataset.

Summary and Key Insight

The literature consistently demonstrates that:

- CAN vulnerabilities are structural and universal
- All major attack types are experimentally validated
- Cryptographic solutions are not scalable for legacy fleets
- Existing IDS research focuses on detection only, not full defense

Literature Synthesis

Across more than two decades of research, a consistent consensus has emerged regarding CAN security. First, the vulnerabilities of the CAN protocol are structural and inherent to its original design assumptions rather than implementation flaws. The absence of authentication, encryption, and access control is fundamental to the protocol architecture itself.

Second, extensive experimental studies have demonstrated the feasibility of a wide range of attacks, including spoofing, replay, denial-of-service, and ECU bus-off exploitation, under both laboratory and real-vehicle conditions. These findings confirm that CAN-based systems can be compromised using relatively low-cost and widely available tools.

Third, cryptographic enhancements proposed in the literature provide theoretical security improvements; however, their practical deployment remains constrained by strict real-time requirements, limited frame payload size, and compatibility issues with legacy automotive electronic control units.

Finally, due to these constraints, research in automotive cybersecurity has progressively shifted away from protocol modification toward monitoring-based defense strategies. In particular, intrusion detection systems leveraging statistical learning, machine learning, and deep learning have become the dominant research direction, as they operate externally to the protocol stack and preserve backward compatibility.

Overall, the literature establishes a clear transition from prevention-oriented protocol design toward detection-centric security mechanisms, forming the foundation for modern CAN bus cybersecurity research.

Systematic Review Methodology

A structured literature screening process, inspired by PRISMA principles, was followed to identify representative CAN bus intrusion detection works and enable a consistent capability comparison across studies. The methodology emphasizes architectural capability analysis rather than procedural review statistics.

Research Questions

The review is guided by four research questions targeting architectural trends, capability coverage, dataset usage, and missing functional dimensions in CAN IDS research.

Research Questions

ID	Research Question
RQ1	What ML/DL architectures have been applied to CAN IDS between 2020 and 2025?
RQ2	To what extent do existing approaches incorporate explainability, autonomous response, and SOC integration?
RQ3	What datasets are used for evaluation, and is multi-dataset validation commonly adopted?

ID	Research Question
RQ4	Which capability dimensions remain systematically unaddressed in current literature?

Search Strategy

A structured search was conducted across IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect targeting peer-reviewed publications from 2020 to 2025. The Boolean query combined terms related to CAN bus, intrusion detection, and machine/deep learning models to ensure coverage of relevant IDS research.

Studies were included if they proposed or evaluated ML/DL-based CAN IDS solutions with empirical validation. Survey papers, non-automotive CAN applications, preprints, and non-peer-reviewed works were excluded to maintain comparability.

Capability Mapping Protocol

Each selected study was evaluated across eleven capability dimensions including detection approach, deep learning usage, multi-attack handling, explainability (XAI), autonomous response, SOC integration, and dataset diversity.

A binary coding scheme was used where ✓ indicates explicit implementation supported by evidence and – indicates absence or non-documentation. This mapping enables structured comparison of functional maturity across the literature.

Methodological Limitations

The review is subject to typical literature study limitations including publication bias, database coverage constraints, language restrictions, and reliance on explicitly reported system capabilities. Industry implementations and non-indexed technical reports may not be fully represented.

Threat Model and Attack Taxonomy

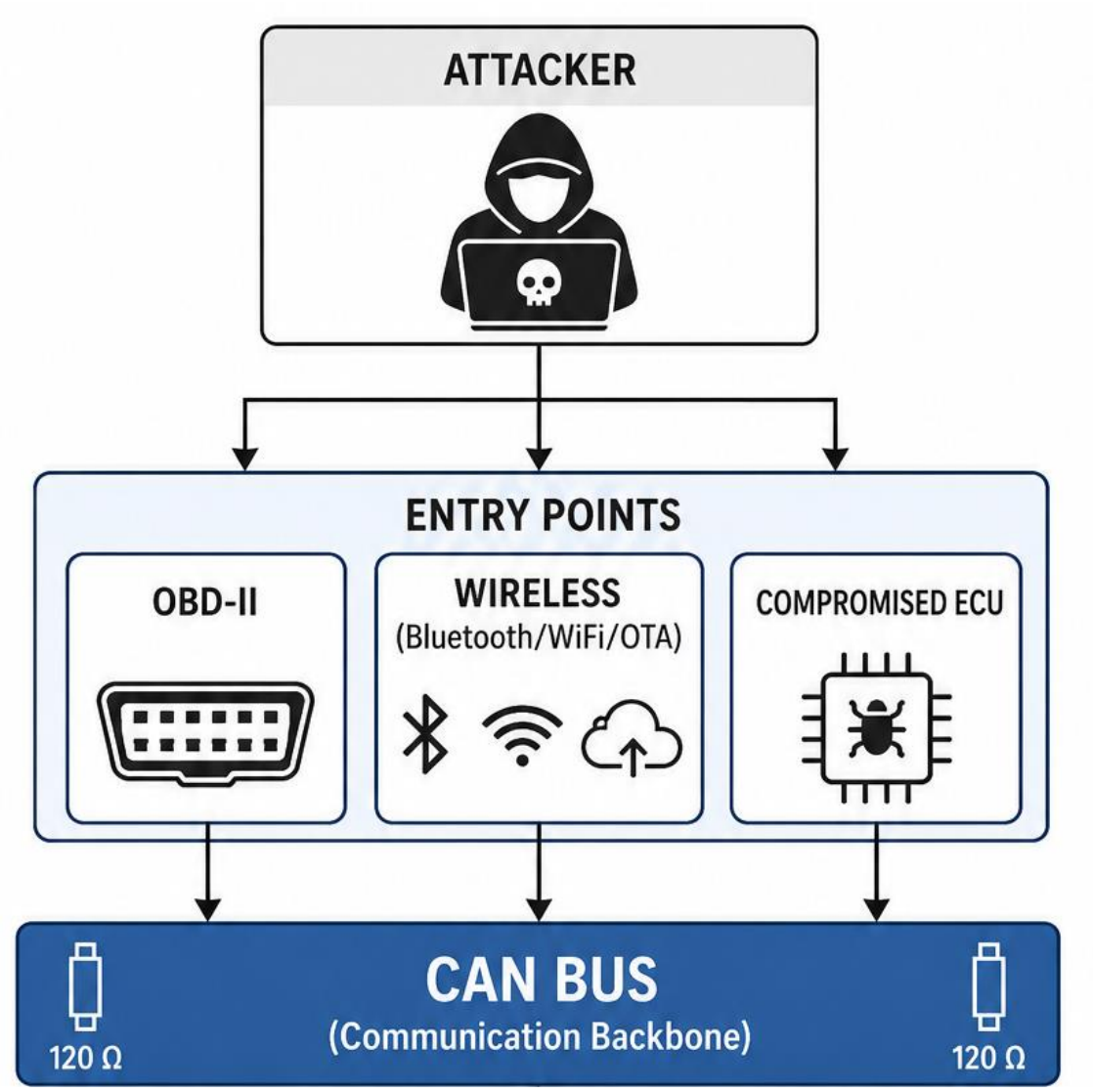
This study adopts a widely used and well-established CAN-bus threat model in automotive cybersecurity research [3], [27], [29]. The model assumes that the adversary has successfully obtained the ability to transmit messages on the in-vehicle CAN network. This access is realistic in modern connected vehicles due to multiple attack surfaces, including compromised electronic control units (ECUs), such as infotainment or telematics systems, physical access through the OBD-II diagnostic port, malicious aftermarket devices (e.g., dongles), wireless interfaces such as Bluetooth or tire-pressure monitoring systems (TPMS), and cloud-based or OTA update mechanisms.

Once this access is achieved, the adversary operates with two fundamental capabilities. First, they can passively monitor (sniff) all CAN traffic due to the broadcast nature of the protocol. Second, they can actively inject arbitrary CAN frames with forged identifiers and manipulated payloads. Since the CAN protocol lacks authentication, encryption, and message integrity verification, even a single compromised node can affect the behavior of multiple safety-critical ECUs. This makes the system highly vulnerable to message-level attacks that do not require full system compromise.

Adversary Capability Model

To systematically analyze threat severity, we define a three-tier adversary model commonly

used in CAN security literature. Each tier represents increasing levels of access, capability, and impact on the in-vehicle network.



CAN Threat Model Architecture

Adversary Capability Model

Tier	Access Level	Capability Description
Tier-1	Physical access	The attacker has direct physical access to the vehicle network via the OBD-II port or direct wiring. This allows immediate message injection and traffic observation without needing to compromise any ECU.
Tier-2	Compromised ECU	The attacker compromises an internal vehicle component such as infotainment, Bluetooth module, or telematics control unit. This enables indirect access to the CAN bus through legitimate communication pathways.

Tier	Access Level	Capability Description
Tier-3	Remote attacker	The attacker exploits external connectivity such as cellular, Wi-Fi, or OTA update mechanisms, allowing remote entry into the vehicle network without physical presence.

This classification shows that modern cars are becoming networked cyber physical systems, and that the same influence an adversary can have over the system can be done remotely as well.

CAN Bus Attack Taxonomy

Attack Type	Description
DoS / Bus Flooding	During such an attack, the attacker sends a massive amount of high-priority messages with an ID number of 0x000. Because of the priority-based arbitration nature of CAN, such high-priority messages take over the bus and starve the other ECU devices from communicating.
Fuzzy Injection	These CAN frames have unpredictably generated or malformed IDs and payload data. Such attacks are typically carried out for testing the performance of the ECU, generating faults in the system, or exploiting any flaws in message handling and processing.
Spoofing	In a spoofing attack, the attacker spoofs an authentic ECU by using its identifier but manipulating its payload data. In one such attack, the attacker could spoof critical vehicle parameters like vehicle speed or RPM of the engine, thereby resulting in erroneous actions by the authentic ECU.
Replay Attack	The messages that have been earlier transmitted are now re-transmitted at a different point in time, regardless of any system-specific environment. This would cause the ECU to receive stale sensor data as if it were current..
Masquerade Attack	This is a sophisticated method of spoofing that involves shutting down or suppressing the legitimate ECU using bus-off attacks or error injection techniques. This is followed by taking over the identity of the suppressed ECU to make detection much harder.
Suspension Attack	In this scenario, the attacker would ensure that the legitimate ECU does not send messages through the CAN Bus. This would mean that some crucial data related to sensors or control systems is unavailable.

These attack types capture both low-level communication manipulation and high-impact system disruption behaviors.

Taxonomy of CAN IDS Approaches

Based on the systematic review, CAN IDS approaches are classified into five architectural categories.

Statistical and Frequency-Based IDS

Early approaches used entropy analysis, Hamming distance, and inter-arrival time statistics. Marchetti and Stabili [14]-[15] proposed entropy-based and sequence-based anomaly detectors. These methods are computationally light and partially interpretable but rely on stationarity assumptions that fail under driver-behavior drift.

Machine Learning-Based IDS

Tree-based classifiers using hand-engineered features represent the first ML generation. Avatefipour et al. [13] applied a modified one-class SVM. Taylor et al. [16] applied LSTM to anomaly detection. These approaches improved detection accuracy but remained single-dataset, black-box systems.

Deep Learning-Based IDS

Deep learning dominates post-2020 literature. Song et al. [21] applied deep CNN to time-windowed ID matrices. Hossain et al. [17] introduced LSTM-based detection. Mehedi et al. [19] employed BiLSTM for improved temporal context. Hanselmann et al. [26] introduced CANet, an LSTM-autoencoder for unsupervised detection.

Hybrid and Ensemble IDS

Several works combine statistical with ML approaches [23], [25]. Javed et al. [23] proposed CANintelliIDS combining CNN with attention-based GRU. While hybrids improve robustness, they reproduce the alert-only paradigm.

Specification and Voltage-Based IDS

Physical-layer fingerprinting — voltage signatures, clock skew — identifies transmitting ECUs [36]-[37]. These methods provide strong identity binding but require specialized hardware and lack integration with XAI or autonomous response.

Critical Review of Existing CAN IDS

Key Findings from the Literature

Despite high reported accuracy (>99% on benchmarks), every reviewed system shares the same architectural limitation: the model emits a class label or anomaly score, and the pipeline ends there. No autonomous response, no explanation, no SOC integration, and no whitelisting are implemented.

Specific Limitations by Approach

Limitations of Existing Approaches

Approach	Limitations
Statistical	Stationarity assumption, fails on complex attacks
ML-based	Single-dataset, feature engineering dependent
DL-based	Black-box, no explainability, alert-only
Hybrid	Reproduces alert-only paradigm
Voltage-based	Hardware-dependent, no XAI or SOC

The Black-Box Problem in ML/DL IDS

Deep neural classifiers are universal approximators with millions of parameters whose decisions are functionally opaque. The automotive application brings four practical issues with it: Within the auto domain the lack of transparency causes four practical issues.

Regulatory Compliance

ISO/SAE 21434 mandates traceability of security decisions throughout the vehicle lifecycle [12]. The EU AI Act classifies safety-critical automotive AI as high-risk, requiring human oversight and explainability [41]. Black-box IDS cannot satisfy these requirements.

Analyst Trust

When a security analyst receives an alert, the absence of per-alert explanation forces binary choice between blind trust and blind dismissal. Operational research shows this is precisely where alert fatigue emerges [38].

Debugging and Forensics

When a model misclassifies, engineers cannot determine whether failure stems from a feature, temporal pattern, or labeling artifact. Explainability is a prerequisite for responsible deployment of safety-critical AI [39].

Adversarial Robustness

Black-box models conceal decision boundaries from defenders. CAN IDS classifiers are susceptible to adversarial frame perturbations [40]. Without explanations, defenders cannot harden fragile decision regions.

Explainable AI (XAI) for Intrusion Detection

SHAP (SHapley Additive exPlanations)

SHAP [42] derives feature attributions from coalitional game theory, assigning each feature a Shapley value quantifying its marginal contribution to the prediction. SHAP satisfies local accuracy, missingness, and consistency. In network IDS contexts, SHAP has been applied to intrusion detection benchmarks [43], but its application to CAN IDS remains limited.

LIME (Local Interpretable Model-agnostic Explanations)

LIME [44] explains individual predictions by training an interpretable surrogate on perturbations of the instance, weighted by proximity. LIME produces human-readable rules such as "frame_rate > 0.85 increased DoS probability by 0.42."

Counterfactual Explanations

Counterfactuals [46] answer: "What minimal change to the input would produce a different decision?" They identify feature interventions a defender could perform to render a frame benign.

Current State of XAI in CAN IDS

Our review found no published CAN IDS that combines all three methods to produce per-alert explanations for SOC analysts. Two recent papers [49]-[50] apply SHAP post-hoc but treat XAI as an offline analysis artifact, not an integrated pipeline component.

Security Operations Center (SOC) Integration

SOC Principles

The Security Operations Center (SOC) is an organizational and technical structure that gathers, correlates, analyses and reacts to security-related telemetry from the systems it monitors [38, 51]. It brings together Security Information and Event Management (SIEM), Security Orchestration, Automation and Response (SOAR) and threat intelligence systems to facilitate coordinated incident response.

To automotive cybersecurity this concept is extended to a Vehicular SOC (V-SOC), which provides continuous monitoring and analysis

of the telemetry data in real time from connected vehicle fleets [52]. The vehicles are distributed sensing nodes that collectively provide 'centralized' security intelligence.

- Regulatory Requirements

UNECE WP.29 R155 reiterates the need for SOC level monitoring, including the ability to detect, monitor and respond to cybersecurity incidents during the vehicle lifecycle, for manufacturers who are required to establish a Cybersecurity Management System (CSMS) [12]. To achieve these, the V-SOC will be the operational implementation of this requirement, providing fleet-wide view and coordinated action.

- Current Integration Gap

Although the intrusion detection system capabilities of CAN-bus have been improved, to date no work in the literature has combined outputs of the IDS with a fully operational V-SOC. Existing IDS solutions invariably end at the classification level, with either anomaly scores or attack labels, and without system level integration. This leaves a key disconnect between detection and response in automotive security systems. A V-SOC integration layer should contain:

- Real-time alert streaming,
- er-alert explainability integration (XAI-based interpretation) [42], [44],
- Risk heatmaps and time-series analysis, and
- Policy-based response mechanisms (for example, CAN ID filtering, ECU isolation), and
- An organized logging system for post event analysis.

Lack of these capabilities signifies that IDS research is still centered around detection accuracy and not full deployment of the cybersecurity system.

Datasets and Evaluation Challenges

- Available Public Datasets

There exist several publicly accessible data sets for IDS research on CAN bus intrusions, with different vehicles used as test subjects and different methods of attack implemented.

Public CAN Bus Datasets

Dataset	Year	Vehicle	Key Characteristic
HCRL Car-Hacking [27]	2018	Hyundai Sonata	Most widely used benchmark dataset
OTIDS [53]	2017	Kia Soul	Includes remote frame injection attacks
SynCAN [26]	2020	Synthetic	Algorithmically generated CAN traffic
ROAD [29]	2024	Ford Expedition	High realism with real driving conditions
CAN-MIRGU [54]	2024	Moving vehicle	Real-world driving scenario dataset
CAN-Train/Test [55]	2023	Hyundai platform	Standardized train-test split design

- Single-Dataset Overfitting

One of the primary limitations in previous research on CAN IDS is the high dependence on testing based on a single dataset, which is usually the HCRL Car-Hacking dataset. Although previous literature claims an accuracy rate greater than 99% for IDS models tested on HCRL, their accuracy does not translate to different realistic scenarios.

Studies have shown [29], [26] that IDS models developed using HCRL show poor accuracy rates (around 70-80%) when tested using realistic datasets such as ROAD and SynCAN. This decline in accuracy is mainly due to factors such as differences in vehicle design, CAN ID scheduling, traffic, and attacks. These results show that current IDS models exhibit significant overfitting.

Multi-Dataset Training Gap

The use of multi-dataset training, which utilizes joint learning from datasets like HCRL, OTIDS, SynCAN, ROAD, and CAN-MIRGU using the same label set, has been suggested as a way forward to tackle generalization issues. Nonetheless, our comprehensive review suggests that no CAN IDS framework makes use of the multi-dataset training technique.

Evaluation Metrics

Assessing the effectiveness of CAN IDS should not be limited to accuracy, which can be highly deceptive in the case of extremely unbalanced data sets, where the proportion of normal data exceeds 95%. Evaluating a CAN IDS involves assessing its efficacy and feasibility.

When dealing with safety-critical automotive applications, low latency detection becomes important. Denial of Service (DoS) attacks can overwhelm the CAN network in under 10 ms; hence, detection and action should take place within strict real-time parameters, preferably less than 50 ms.

Conversely, the false positive rate (FPR) becomes important when considering fleets of cars. A FPR of even 0.1% will produce thousands of false positives per hour in large-scale deployments.

Therefore, a comprehensive evaluation framework must report:

- precision, recall, and F1-score (per-class),
- false positive rate (FPR),
- detection latency (mean and 99th percentile), and
- robustness under class imbalance conditions.

Comparative Capability Matrix

Table X displays a systematic analysis of the capability of CAN intrusion detection systems (2020–2025), considering eleven aspects. All of the papers are analyzed based on concrete implementation evidence that is available in the literature, where ✓ denotes that the capability exists and – means it does not exist. Note that the inclusion criteria for Table X are the publication of an article. The results from this comparative analysis of capabilities highlight the necessity of developing a framework for CAN IDS, discussed in the following section.

Work / Year	Detect	DL	Multi-Atk	XAI	SHAP	LIME	Auto Resp	Rate Lim	ECU Off	SOC	Multi-DS
Song et al. 2020 [21]	✓	✓	✓	–	–	–	–	–	–	–	–
Hossain et al. 2020 [17]	✓	✓	✓	–	–	–	–	–	–	–	–
Hanselmann	✓	✓	✓	–	–	–	–	–	–	–	–

Work / Year	Detect	DL	Multi-Atk	XAI	SHAP	LIME	Auto Resp	Rate Lim	ECU Off	SOC	Multi-DS
et al. 2020 [26]											
Mehedi et al. 2021 [19]	✓	✓	✓	–	–	–	–	–	–	–	–
Verma et al. 2024 [29]	✓	✓	✓	–	–	–	–	–	–	–	✓
Alkhatib et al. 2023 [49]	✓	✓	✓	✓	✓	–	–	–	–	–	–
Lampe & Meng 2023 [55]	✓	✓	✓	–	–	–	–	–	–	–	✓
M. H. Shahriar et al. 2024 [50]	✓	✓	✓	✓	✓	–	–	–	–	–	–
S. Jeong, S. Lee et al. 2024 [22]	✓	✓	✓	–	–	–	–	–	–	–	–
J. Lee et al. 2025 [56]	✓	✓	✓	✓	–	✓	–	–	–	–	–

✓ = Present, – = Absent, DL= Deep Learning, Multi-DS = Multi-Dataset Training

Key observations from the matrix:

1. All reviewed works allow detection, yet only four try any XAI approach
2. None of the reviewed works consider autonomous response (ratelimiting, dropping frames, ECUBus-off)
3. None of the reviewed works use a SOC dashboard
4. Training on multiple datasets is found in only two papers [29], [55]

Major Research Gaps in CAN Bus Security

It is evident from the above discussion and the quantitative analysis in Table X that the CAN IDS research community follows a certain trend: Previous studies analyze the intrusion detection problem as a classification problem based on accuracy measures instead of a security architecture deployable in real-world vehicles [4], [27], [29].

The mentioned gaps are interlinked and symptomatic to this architectural flaw. They fall into five main categories.

Gap 1: Detection-Only, Alert-Only Paradigm

All reviewed CAN IDS end at the classifier output. The system generates the anomaly value or classifies the message as being an attack. This is where all papers cease to contribute to the research. None of them combines IDS with any form of autonomous defense, such as ID filtering, rate limiting,

frame dropping, or ECU isolation [14], [19], [21], [26].

The reason for this is clearly illustrated by the consistently blank *Auto Response*, *Rate Limiting*, and *ECU Off* columns shown in Table X. IDSs are always viewed as testing devices instead of a means of protection against attacks, which leads to the discrepancy between scholarly products and actual automobile security needs.

Gap 2: Black-Box Decisions Without Explainability

There are no per-alert explanations of the model output based on SHAP, LIME, or counterfactual explanations in CAN IDS workflows. Some few researchers have employed techniques such as XAI after training models, but there is no integration of explanation within the inference process for use [42], [43], [44], [49], [50].

It goes against the traceability criteria of ISO/SAE 21434 and AI regulation expectations for transparency in AI safety critical applications. There is no way to audit or rely on an alert that does not include an explanation.

Gap 3: Absence of SOC Integration

Although mandatory regulation exists in terms of monitoring and mitigating cybersecurity threats at the fleet level, there is currently no CAN IDS methodology published that is built to support integration into the vehicle-specific SOC infrastructure [12], [38], [51], [52]. This includes real-time alerts, explanations, fleet-wide correlation capabilities, manual control options, and forensics tracking.

Because of this, current models of IDS can not be seamlessly integrated into a functional management system.

Gap 4: Single-Dataset Overfitting and Poor Generalization

IDS model training and evaluation are carried out using one dataset only, and this usually happens using the HCRL Car Hacking dataset [27]. Cross-dataset evaluation is not common, while multi-dataset training is practically unheard of [26], [29].

Research has indicated that IDS models trained and tested using a particular dataset have very poor performances when tested against other datasets because of differences in the way the vehicles operate, ID scheduling, and attacks on the network.

Gap 5: Fragmented Literature, No Unified Architecture

There have been various techniques developed by individual researchers such as XAI models, voltage fingerprinting, rate limiting techniques, and behavioral analysis that operate independently [36], [37], [42], [46]. However, there is no prior literature that combines all of these to form a comprehensive module that can be implemented.

There is no literature that considers these aspects in one study as part of an integrated CAN security framework.

Directions for Future Research on XAI and SOC-Integrated CAN IDS

Based on the research gaps mentioned above and on the proposed architecture, this section attempts to present a systematic research agenda for next-generation CAN intrusion detection systems. The objective is to move

from traditional detection techniques to the development of intelligent cybersecurity solutions.

A. Integration of XAI into CAN IDS Pipelines

Future work must progress beyond post hoc explainability approaches and incorporate techniques like SHAP, LIME, and counterfactual reasoning into the prediction process itself. The goal is to deliver per-alarm explanations that are relevant for SOC operators.

Some key issues to address include delivering sub-50 ms inference time, delivering accurate explanations for CAN-related features, and developing comprehensible user interfaces for automotive security operators.

B. Autonomous Response Mechanisms

An important area of research would be the design of real-time response solutions that go beyond just detection. This would include:

- (i) dynamic CAN ID whitelisting/blacklisting.
- (ii) adaptive rate limiting using sliding window models.
- (iii) frame filtering or destruction at gateway level.
- (iv) controlled ECU isolation through bus-off mechanisms.

Such mechanisms must be formally verified against functional safety standards such as ISO 26262 to ensure that security actions do not compromise vehicle safety.

C. SOC Integration for Fleet-Level Security

Future CAN IDS architectures should be integrated with vehicular SOC (V-SOC) systems to enable fleet-wide monitoring and response. Required capabilities include real-time alert streaming, per-alert XAI explanation visualization, cross-vehicle threat correlation, manual override controls, and forensic-grade logging aligned with regulatory requirements such as UNECE WP.29 R155.

D. Multi-Dataset Training and Benchmark Standardization

To address generalization limitations, future work should adopt multi-dataset training strategies combining datasets such as HCRL Car-Hacking, OTIDS, SynCAN, ROAD, CAN-MIRGU, and CAN-Train under unified labeling schemas.

Standard evaluation protocols should include: per-class precision, recall, F1-score, detection latency (mean and 99th percentile), false positive rate at fleet scale, and cross-dataset transfer performance.

E. Federated and Privacy-Preserving Learning

Federated learning offers a scalable solution for privacy-preserving fleet intelligence. All vehicles learning independently but sharing encrypted models without sharing raw CAN data. Some of the key issues to address would be data distribution being non-IID, different hardware for ECUs, and limited connectivity.

F. Adversarial Robustness and Secure Explainability

IDSs in the future have to be resilient to adversarial attacks on both classifiers and explanation models. This calls for research into ensembles, robustness certification, and security in explanations to counteract manipulation of SHAP and LIME outputs.

G. Hardware-Software Co-Design for Real-Time Deployment

Deployment within production cars must adhere strictly to computational restrictions.

This will involve concentrating on compressing the models, quantizing them, and accelerating them with automotive-grade computing devices like the Infineon AURIX and ARM-based ECUs. The goal is to attain determinism of line-rate detection and reaction.

H. Standardization and Certification Pathways

Lastly, it should be noted that any future development needs to make the transition from algorithms to compliance. Compliance with ISO/SAE 21434, ISO 26262, and AUTOSAR cybersecurity standards is a must. It will be necessary to standardize the format of the outputs produced by the system.

Limitations of This Review

Several aspects of the methodology used in this systematic review require consideration.

- **Publication bias:** The literature reviewed is likely biased towards positive results since articles describing high performance in detection are more often published than those that report negative results. This might lead to an overstatement regarding the maturity of CAN IDS techniques.
- **Database coverage:** Four major academic databases (IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect) were considered. Other relevant articles published in venues that are not indexed by these databases, industry reports, or patents may be missed.
- **Language bias:** Only articles written in English language were considered. There may be some studies conducted outside English-speaking countries which could also be relevant.
- **Extraction bias:** The capability of each implementation is based on explicit data extracted from the original papers. Not all implementations are documented sufficiently, while others may exist but do not have a detailed description in the paper's main text.
- **Temporal scope:** This systematic review considers papers published during 2020-2025. Older foundational papers are referenced for context but not analyzed under the same methodology.

Conclusion

This paper has conducted a systematic review of current CAN intrusion detection research literature. By analyzing through various capability perspectives, it can be found that existing literature has mostly been converging towards an approach that detects only, is black-box based, and works only with a single data set. Although highly accurate results have been achieved using these algorithms against benchmark data sets, yet their explainability, autonomy, SOC integration, and generalization to other vehicles is lacking.

A total of five major challenges in the field were formulated in the following manner:

- Detection Only Approach.
- Black Box Decision Models Lacking Explainability.
- No Integration with SOC.
- Overfitting and lack of generalization among different data sets.
- Lack of architectural standards among the literature.

The future work of this paper calls for an integrated approach that includes explainability, autonomous response mechanism, multiple datasets, and SOC integration under one architecture.

Overall, there is a requirement for transitioning from highly accurate but unexplainable and non-operational CAN Intrusion detection models to automotive cybersecurity systems.

REFERENCES

1. M. Bozdal, M. Samie, S. Aslam, and I. Jennions, "Evaluation of CAN bus security challenges," *Sensors*, vol. 20, no. 8, p. 2364, Apr. 2020.
2. S. Nie, L. Liu, and Y. Du, "Free-fall: Hacking Tesla from wireless to CAN bus," in *Proc. Black Hat USA*, Las Vegas, NV, USA, Jul. 2017, pp. 1–16.
3. "Road vehicles — Controller area network (CAN) — Part 1: Data link layer and physical signalling," ISO 11898-1:2015, International Organization for Standardization, Geneva, Switzerland, 2015.
4. S. Checkoway et al., "Comprehensive experimental analyses of automotive attack surfaces," in *Proc. USENIX Security Symp.*, San Francisco, CA, USA, Aug. 2011, pp. 1–16.
5. C. Miller and C. Valasek, "A survey of remote automotive attack surfaces," in *Proc. Black Hat USA*, Las Vegas, NV, USA, Aug. 2014, pp. 1–94.
6. K. Koscher et al., "Experimental security analysis of a modern automobile," in *Proc. IEEE Symp. Security Privacy*, Oakland, CA, USA, May 2010, pp. 447–462.
7. P. Kleberger, T. Olovsson, and E. Jonsson, "Security aspects of the in-vehicle network in the connected car," in *Proc. IEEE Intell. Veh. Symp.*, Baden-Baden, Germany, Jun. 2011, pp. 528–533.
8. Greenberg, "Hackers remotely kill a Jeep on the highway—with me in it," *Wired*, Jul. 2015.
9. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," in *Proc. Black Hat USA*, Las Vegas, NV, USA, Aug. 2015, pp. 1–91.
10. T. Hoppe, S. Kiltz, and J. Dittmann, "Security threats to automotive CAN networks — Practical examples and selected short-term countermeasures," *Reliab. Eng. Syst. Saf.*, vol. 96, no. 1, pp. 11–25, Jan. 2011.
11. S. Woo, H. J. Jo, and D. H. Lee, "A practical wireless attack on the connected car and security protocol for in-vehicle CAN," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 993–1006, Apr. 2015.
12. "Road vehicles — Cybersecurity engineering," ISO/SAE 21434:2021, International Organization for Standardization, Geneva, Switzerland, 2021.
13. O. Avatefipour et al., "An intelligent secured framework for cyberattack detection in electric vehicles' CAN bus using machine learning," *IEEE Access*, vol. 7, pp. 127580–127592, 2019.
14. M. Marchetti and D. Stabili, "Anomaly detection of CAN bus messages through analysis of ID sequences," in *Proc. IEEE Intell. Veh. Symp.*, Los Angeles, CA, USA, Jun. 2017, pp. 1577–1583.
15. D. Stabili, M. Marchetti, and M. Colajanni, "Detecting attacks to internal vehicle networks through Hamming distance," in *Proc. AEIT Int. Annu. Conf.*, Cagliari, Italy, Sep. 2017, pp. 1–6.
16. Taylor, S. Leblanc, and N. Japkowicz, "Anomaly detection in automobile control network data with long short-term memory

networks," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, Montreal, QC, Canada, Oct. 2016, pp. 130–139.

17. M. D. Hossain, H. Inoue, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based intrusion detection system for in-vehicle CAN bus communications," *IEEE Access*, vol. 8, pp. 185489–185502, 2020.
18. M. D. Hossain et al., "An effective in-vehicle CAN bus intrusion detection system using GRU," in *Proc. IEEE BCD*, 2020, pp. 1–6.
19. S. T. Mehedi, A. Anwar, Z. Rahman, and K. Ahmed, "Deep transfer learning based intrusion detection system for electric vehicular networks," *Sensors*, vol. 21, no. 14, p. 4736, Jul. 2021.
20. Z. Khan, M. Chowdhury, M. Islam, C.-Y. Huang, and M. Rahman, "Long short-term memory neural network-based attack detection model for in-vehicle network security," *IEEE Sens. Lett.*, vol. 4, no. 6, pp. 1–4, Jun. 2020.
21. H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Veh. Commun.*, vol. 21, p. 100198, Jan. 2020.
22. S. Jeong, S. Lee, H. Lee, and H. K. Kim, "X-CANIDS: Signal-Aware Explainable Intrusion Detection System for Controller Area Network-Based In-Vehicle Network," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 3230–3246, Mar. 2024. DOI: 10.1109/TVT.2023.3327275
23. R. Javed, S. Ur Rehman, M. U. Khan, M. Alazab, and T. R. G., "CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1456–1466, Apr.–Jun. 2021.
24. S. Tariq, S. Lee, and S. S. Woo, "CANTransfer: Transfer learning based intrusion detection on a controller area network using convolutional LSTM network," in *Proc. ACM SAC*, Brno, Czech Republic, Mar.–Apr. 2020, pp. 1048–1055.
25. W. Lo, H. He, H. Li, A. Patel, L. Pan, and J. Jiang, "A hybrid deep learning based intrusion detection system using spatial-temporal representation of in-vehicle network traffic," *Veh. Commun.*, vol. 35, p. 100471, Jun. 2022. DOI: 10.1016/j.vehcom.2022.100471
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. 58194–58205, 2020.
27. H. Lee, S. H. Jeong, and H. K. Kim, "OTIDS: A novel intrusion detection system for in-vehicle network by using remote frame," in *Proc. IEEE PST*, Toronto, ON, Canada, Dec. 2017, pp. 57–66.
28. G. Dupont, J. den Hartog, S. Etalle, and A. Lekidis, "A survey of network intrusion detection systems for controller area network," in *Proc. IEEE ICVES*, Cairo, Egypt, Sep. 2019, pp. 1–6.
29. M. E. Verma et al., "Addressing the lack of comparability and testing in CAN intrusion detection research: A comprehensive guide to CAN IDS data and introduction of the ROAD dataset," *PLOS ONE*, vol. 19, no. 1, p. e0296879, Jan. 2024.

30. C. Ling and D. Feng, "An algorithm for detection of malicious messages on CAN buses," in *Proc. CCISP*, 2012, pp. 1–5.
31. K.-T. Cho and K. G. Shin, "Fingerprinting electronic control units for vehicle intrusion detection," in *Proc. USENIX Security Symp.*, Austin, TX, USA, Aug. 2016, pp. 911–927.
32. K.-T. Cho and K. G. Shin, "Error handling of in-vehicle networks makes them vulnerable," in *Proc. ACM CCS*, Vienna, Austria, Oct. 2016, pp. 1044–1055.
33. P. Carsten, T. R. Andel, M. Yampolskiy, and J. T. McDonald, "In-vehicle networks: Attacks, vulnerabilities, and proposed solutions," in *Proc. ACM CISR*, Oak Ridge, TN, USA, Apr. 2015, pp. 1–8.
34. B. Groza and P.-S. Murvay, "Security solutions for the controller area network," *IEEE Veh. Technol. Mag.*, vol. 13, no. 1, pp. 36–44, Mar. 2018.
35. "Specification of Secure Onboard Communication," AUTOSAR, Release 4.4.0, AUTOSAR Consortium, 2019.
36. M. Foruhandeh, Y. Man, R. Gerdes, M. Li, and T. Chantem, "SIMPLE: Single-frame based physical layer identification for intrusion detection and prevention on in-vehicle networks," in *Proc. ACSAC*, San Juan, PR, USA, Dec. 2019, pp. 229–244.
37. W. Choi, K. Joo, H. J. Jo, M. C. Park, and D. H. Lee, "VoltageIDS: Low-level communication characteristics for automotive intrusion detection system," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2114–2129, Aug. 2018.
38. C. Zimmerman, "Ten Strategies of a World-Class Cybersecurity Operations Center," MITRE Corporation, McLean, VA, USA, Tech. Rep., 2014.
39. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
40. F. Pollicino, D. Stabili, L. Ferretti, and M. Marchetti, "An experimental analysis of ECML-PKDD 2021 evasion attacks on automotive intrusion detection systems," *Veh. Commun.*, vol. 36, p. 100487, Aug. 2022.
41. "Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, 2024.
42. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
43. M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.
44. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM KDD*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.
45. L. Yang, A. Moubayed, I. Hamieh, and A. Shami, "Tree-based intelligent intrusion detection system in Internet of Vehicles," in *Proc. IEEE GLOBECOM*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.
46. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening

the black box: Automated decisions and the GDPR," *Harv. J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

47. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. ACM FAccT*, Barcelona, Spain, Jan. 2020, pp. 607–617.
48. S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *Proc. PPSN*, Leiden, Netherlands, Sep. 2020, pp. 448–469.
49. 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 *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. (AICCSA)*, Abu Dhabi, UAE, Dec. 2022, pp. 1–8. DOI: 10.1109/AICCSA56895.2022.10017800
50. M. H. Shahriar, Y. Xiao, P. Moriano, W. Lou, and Y. T. Hou, "CANShield: Deep-Learning-Based Intrusion Detection Framework for Controller Area Networks at the Signal Level," *IEEE Internet of Things J.*, vol. 10, no. 24, pp. 22111–22127, Dec. 2023. DOI: 10.1109/JIOT.2023.3302271
51. V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in *Proc. EISIC*, Athens, Greece, Sep. 2017, pp. 91–98.
52. M. Strandberg, T. Olovsson, and B. Nilsson, "Securing the connected car: A security-enhancement methodology," *IEEE Veh. Technol. Mag.*, vol. 13, no. 1, pp. 56–65, Mar. 2018.
53. Hacking and Countermeasure Research Lab, "OTIDS Dataset," 2017. [Online]. Available: https://ocslab.hksecurity.net/Datasets/OTIDS
54. S. Rajapaksha, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, G. Madzudzo, and M. Cheah, "CAN-MIRGU: A comprehensive CAN bus attack dataset from moving vehicles for intrusion detection system evaluation," in *Proc. NDSS Vehicle Security Workshop*, San Diego, CA, USA, Feb. 2024, pp. 1–13.
55. B. Lampe and W. Meng, "can-train-and-test: A curated CAN dataset for automotive intrusion detection," *Comput. Secur.*, vol. 140, p. 103777, 2024. DOI: 10.1016/j.cose.2024.103777
56. J. Lee and J. Rew, "Vision-Language Model-Based Local Interpretable Model-Agnostic Explanations Analysis for Explainable In-Vehicle Controller Area Network Intrusion Detection," *Sensors*, vol. 25, no. 10, p. 3020, May 2025. DOI: 10.3390/s25103020
57. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.
58. S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4734–4746, Aug. 2020.
59. "Architecture enhancements for 5G System (5GS) to support Vehicle-to-Everything (V2X) services," 3GPP, TS 23.287, v18.1.0, Dec. 2023.

60. Infineon Technologies, "AURIX TC4xx Cybersecurity Hardware Reference," Infineon Technologies AG, Neubiberg, Germany, Tech. Rep., 2023.
61. "UNECE WP.29, Regulation No. 155 — Cyber security and cyber security management system," United Nations Economic Commission for Europe, Geneva, Switzerland, 2021.
62. W. Wu et al., "A survey of intrusion detection for in-vehicle networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 919–933, Mar. 2020.