



(RESEARCH ARTICLE)



DO CAN INTRUSION DETECTORS TRAVEL? A CROSS-DATASET AND ZERO-DAY STUDY WITH IDENTIFIER-AGNOSTIC, LOCALLY CALIBRATED ENSEMBLES

Akinwale Obafemi *

School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, Scotland, UK.

* Corresponding Author

ORCID Details

Akinwale Obafemi: <https://orcid.org/0009-0003-6875-1900>

World Journal of Advanced Research and Reviews, 2026, 31(03), 1882–1889

Publication history: Received on 21 August 2026; revised on 24 September 2026; accepted on 26 September 2026

Article DOI: <https://doi.org/10.30574/wjarr.2026.31.3.2509>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Intrusion detection systems (IDS) for the vehicular CAN bus that use Machine Learning are usually evaluated and reported to achieve very high detection accuracy, commonly above 99%, on random splits of a single dataset. We test and reproduce such high detection accuracy for four publicly available CAN IDS benchmark corpora (OTIDS, Car-Hacking, ROAD, SynCAN), including three real-world vehicles and a synthetic generator, and we expose this number as an artifact of the protocol evaluation used so far. We present a detailed evaluation using a highly effective random-forest detector in different settings, i.e., using different corpora for training and testing. Our results show a significant collapse of a detector's mean F1 score from near-perfect in-domain performance to 0.345 when transferred to other corpora. Evidence-based identification of this collapse is a significant contribution of this work, a problem which we subsequently fixed by removing identifier one-hot features and using local statistics that are both identifier-agnostic and are locally calibrated on a short window of unlabelled benign traffic of the target-vehicle. We subsequently achieve a significant increase in mean cross-dataset F1 score without using any labelled attack on the target corpus. To the best of our knowledge, this is the first work that (i) identifies the root-cause of this collapse, namely the feature representation, (ii) presents and repairs this feature-representation to recover much of the lost cross-dataset performance using identifier-agnostic statistics that are locally calibrated on minutes of unlabelled benign target-vehicle traffic, and (iii) reports per-family zero-day recall for four publicly available corpora.

Keywords: CAN Bus, Intrusion Detection, Cross-Dataset Generalization, Zero-Day Attacks, Domain Transfer, Automotive Security

1 INTRODUCTION

The controller area network (CAN) is still the main in-vehicle backbone. Message-injection attacks on the CAN-bus are the best-studied cyber threat to automobiles. Once an attacker has gained access to the vehicle's CAN-bus they can attempt to flood the bus with messages, alter the values of current gauge readings, or even inject completely new messages onto the bus. These new messages can then be sent to safety critical controllers on the vehicle. These attacks

have been demonstrated on real cars [1], [2], [3]. A large body of work now applies machine-learning intrusion detection systems (IDS) to such attacks, and nearly all of it is evaluated on a handful of public datasets, mainly OTIDS [4] and Car-Hacking [5].

Nearly all CAN intrusion detection systems that have been studied employ single-dataset evaluation protocols. For example, a model is trained and tested on a random split of the same data, typically including recordings from the same vehicle and often even the same attacks (typically the same four or five attack recordings). The resulting high accuracy of the detectors (routinely greater than 99%) has recently been challenged in a cross-dataset study [6]. The performance of detection methods on different corpora varies significantly. Systematic audits of current datasets have revealed some problems related to labeling and even artefacts (e.g., [7, 8]), which lead to over-optimistic results within a single dataset. In this work, we first start with reproducing existing results. On our evaluation setup, we achieve an in-domain F1-score of 1.000 for a basic start. However, the naive transfer of these models to other datasets only results in a mean F1-score of 0.345 for a random-forest backbone.

This paper treats the benchmark failure as the object of study and addresses three key questions: 1) How poorly do current CAN IDS models transfer between four different public datasets containing the CAN traffic of three different vehicles plus one synthetic generator, when the attacker families are shared between the datasets? 2) Does the fact that the current representation of the traffic in prior work causes it to memorize the different vehicle-specific CAN identifier(s) for the different nodes in the network mean that such features transfer poorly, and that there are features that do not memorize the identifiers that can be calibrated on a short window of unlabelled traffic from the target vehicle(s)? 3) How do such detectors perform when faced with zero-day (i.e. never before seen) families of attacks that were not used in training?

This paper makes four main contributions: (i) uniform parsing and labeling of four datasets, OTIDS, Car-Hacking, ROAD, and SynCAN; (ii) a set of identifier-agnostic features (inter-arrival times, payload changes, bus rate, and vocabulary statistics) that can be local-normalized and calibrated with nothing more than unlabelled benign traffic from the target vehicle; (iii) a systematic cross-dataset and zero-day test set evaluation of four model types including soft-voting ensembles; (iv) a reporting protocol for performance of CAN IDS methods including per-attack family recall, number of false alarms per hour, and a cross-dataset scorecard. We argue that every CAN IDS paper should report all three. The feature set presented here yields a mean cross-dataset F1 of 0.335 (identifier one-hot representation F1 is 0.147). With no labeled attacks on the target dataset, a soft-voting ensemble of the three tree models here achieves a mean F1 of 0.394 on the twelve possible cross-dataset test sets.

2 RELATED WORK

2.1 CAN Intrusion Detection

Most approaches of In-vehicle IDS have been evaluated on a few existing In-vehicle IDS datasets, like the OTIDS and the Car-Hacking dataset from the Hacking and Countermeasure Research Lab [4], [5]. The Car-Hacking dataset was used for the evaluation of the GIDS approach [5], which uses a GAN-based IDS approach. Mostly the approaches were evaluated on one or two datasets and split randomly into a training- and test set. The threat itself is well documented. Early work showed that in-vehicle networks can be reached and controlled by an attacker [1], [2], [3], and that finding is what started IDS research. Since then the detectors have covered a wide range: lightweight rule- and statistics-based monitors (identifier whitelisting, inter-arrival-time checks [9], entropy monitors [10]), classical learning on frame statistics [11], deep sequence models [12], and GAN-based pipelines such as GIDS [5].

2.2 Datasets and Their Audits

Within the public-domain datasets for intrusion detection in CAN-buses, there are OTIDS [4] and Car-Hacking datasets developed by Hacking and Countermeasure Research Lab (HCRL) [5], ROAD [7] developed by Oak Ridge National Laboratory (ORNL) that includes real fuzzing, fabrication and advanced attacks, among them correlated-signal injection, and simulated masquerade attacks. The other public-domain datasets in the area are, for instance, can-train-and-test datasets [13] that are specially designed for multi-vehicle settings, CAN-MIRGU [14] that includes 36 attack captures in a real moving vehicle, out of which 26 are physically verified for having been injected and the remaining 10 are either masquerade or suspension attacks that have been simulated, and last but not least SynCAN [15] that is a signal-level benchmark implemented in synthetic fashion for in-vehicle networks with CAN-buses. Two audits matter here. can-sleuth [8] examines the capabilities, limitations and performance impacts of six automotive intrusion detection datasets, the HCRL Car-Hacking and Survival Analysis corpora among them, and a 2026 cross-dataset study [6] shows that the same detectors score very differently from one corpus to the next. This paper acts on both critiques, with a representation that fixes the problem they point to and a protocol for reporting it.

2.3 Transfer and Zero-Day Evaluation

Cross-corpus transfer has been studied in many other domains, but in the context of in-vehicle networks it is rare, partly because different vehicles have different identifier vocabularies. In such cases one-hot encodings are broken and, to the best of our knowledge, no prior work has investigated how to isolate the representation bottleneck and ‘fix’ it by using identifier-agnostic statistics that are locally-calibrated to a new corpus. That is the gap this paper fills.

3 METHODOLOGY

3.1 Datasets and Unified Labelling

Table 1 describes the unified data set of 4 corpora. For injection-type attacks, the corresponding ground truth is either found in the documentation of the respective data set (OTIDS: Identifier 0x000 corresponding to DoS floods. Identifier of fuzzed frames not in benign vocabulary) or even explicit flags are available (Car-Hacking: Frames injected into Car-Hacking data set are marked with T. SynCAN: All messages are labeled. ROAD: Injections in captures of ROAD data set are identified by corresponding capture-specific identifiers). Three vehicles of different car makes and one synthetic data generator are selected to cover a range of realism levels.

Table 1: Corpora after unified parsing and labelling

Dataset	Frames	Benign	Attack	IDs	Attack families
OTIDS	3,617,967	3,282,343	335,624	46	benign, dos, fuzzy
Car-Hacking	17,357,681	15,026,164	2,331,517	2048	benign, dos, fuzzy, spoof-gear, spoof-rpm
ROAD	7,871,335	7,776,571	94,764	662	benign, fuzzing, injection
SynCAN	4,400,052	3,150,052	1,250,000	10	benign, continuous, dos, plateau, playback, suppress

3.2 Representations

The default representation for a detector is the representation of per-frame statistics, augmented with an identifier encoding; this is Representation A (Rep-A) in what follows (we used a 64-way hashed one-hot encoding in the experiments). This encoded identifier is used by the detector to ‘learn’ which identifiers are normally present on a vehicle used for training. Our representation, Representation B (Rep-B), removes identifier identity altogether and normalises a frame against the normal behavior of the identifier of that frame on the vehicle where the detector is deployed. The normalization is done using robust z-scores of log inter-arrival time and Hamming distance of payload of the frame. A trailing-rate ratio and a novel-identifier indicator (against the target vocabulary of identifiers) are also included in the Representation. Finally, payload-entropy and some raw byte-level statistics are included as well. Two details matter. The z-scores are taken against per-identifier medians and MADs, and the trailing-rate ratio compares the current rate with the identifier’s own benign rate. All of these calibration quantities come from a short prefix of unlabelled benign traffic on whichever vehicle the detector is deployed on, 15 percent of the benign capture and at most a few minutes, which is what a real commissioning step would look like. No labelled attacks are used at this stage.

3.3 Models

We implement the following models and examine their performance: random forests (RF) and extremely randomised trees (ET), as well as histogram gradient boosting (HGB), and a soft-voting ensemble that combines all the three models. The tree-based ensemble models are particularly suitable for tabular data, can train within minutes on a CPU, and, according to our escalation protocol in Section 5, we moved from a single RF to the ensemble once the single-model cross-dataset results proved weak.

3.4 Protocols and Metrics

For the in-domain experiments we split every corpus in a temporal 60/40 fashion, so the test frames come later in time than the training frames. For cross-dataset experiments we use all benign samples plus all samples of all attack families of a corpus for training and then evaluate on the capped benign sample plus all attack samples of another corpus. For zero-day experiments we remove one attack family from the source corpus’s training data and then measure the recall of this specific family on the other corpora. We report average per-attack-family-recall, as well as alarm rate on benign traffic and false alarms per operating hour on the corresponding benign captures. The per-hour figure comes from the frame rate of each benign capture.

4 EXPERIMENTAL SETUP

We implement the parsing, feature generation, training, and evaluation using Python and the well-established scikit-learn library, and run experiments on a single CPU workstation (running Windows) using Python 3.14 and scikit-learn 1.9. In order to keep runtimes of the evaluations within a reasonable limit, we cap benign training samples at 400k frames and corresponding evaluation benign samples at 200k frames. For attack frames we cap them at 250k frames per family per corpus. In addition, we fix the random seeds for all experiments to ensure reproducibility. All results, feature caches, and the scripts that draw the figures accompany the paper, with the full results stored as one JSON file per condition.

5 RESULTS

5.1 In-Domain Performance and Its Caveats

Table 2 shows the familiar ceiling where the labels are clean. For the OTIDS dataset all models achieve a perfect score of $F1 = 1.000$, as is typical for single-dataset evaluation. The Car-Hacking dataset's performance however is markedly lower: under a temporal, cross-session split (i.e. the benign capture and the attack captures are from different driving sessions) the ensemble $F1$ -score is 0.564, with an alarm rate of 0.000 on the held-out benign-only driving frames (i.e. they were not flagged as injections). In detail, almost all errors reported for Car-Hacking are on benign-labelled frames within the attack recordings, suggesting that the injected frames are not always flagged in the labels. We read this as a labelling artefact, not a model failure. This kind of phenomenon has been documented in public CAN corpora before; dataset audits report unlabelled injections and other artefacts of this kind [7], [8]. Note, that even in the in-domain setting of a standard benchmark there are a lot of caveats, which we allow the reader to inspect in detail for the Car-Hacking dataset in more detail in the per-family numbers we release with the results bundle.

Table 2: In-domain test F1 (temporal split)

Dataset	RF	ET	HGB	Ensemble
OTIDS	1.000	1.000	1.000	1.000
Car-Hacking	0.645	0.962	0.561	0.564
ROAD	0.882	0.885	0.867	0.881
SynCAN	0.454	0.451	0.482	0.469

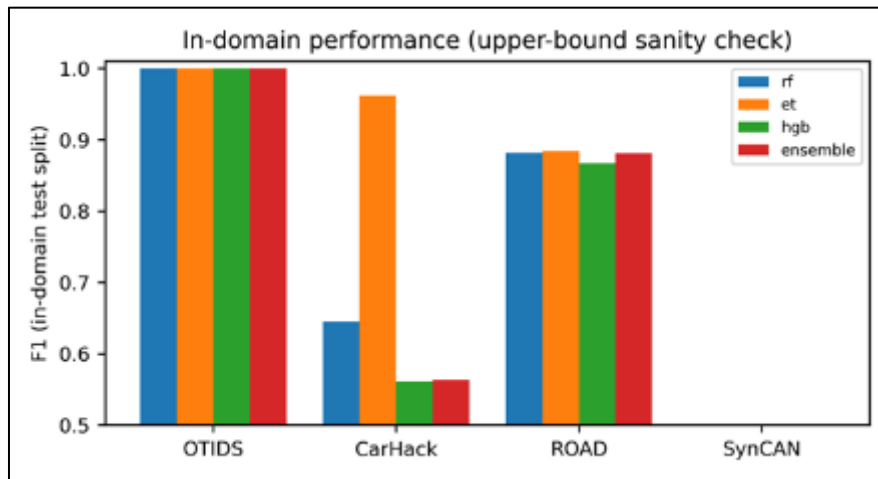


Figure 1: In-domain F1 per model and corpus.

5.2 Cross-Dataset Transfer

The transfer matrix, shown in Figure 2, highlights the collapse in performance between in-domain averages of 0.99 (or greater) for both RF and the ensemble and the corresponding cross-dataset averages of 0.345 (RF) and 0.394 (ensemble). By revealing recall on a per-family basis, it is possible to see which of the tested attacks transfer. The high-rate, simple DoS, detected via physical bus-rate features, transfers. However, attacks that are members of spoofing

families that rely on identifier-specific payload patterns degrade sharply when the identifiers between the in-domain and cross-dataset data differ. The full per-pair, per-model numbers for all tested attacks are available as part of the results bundle.

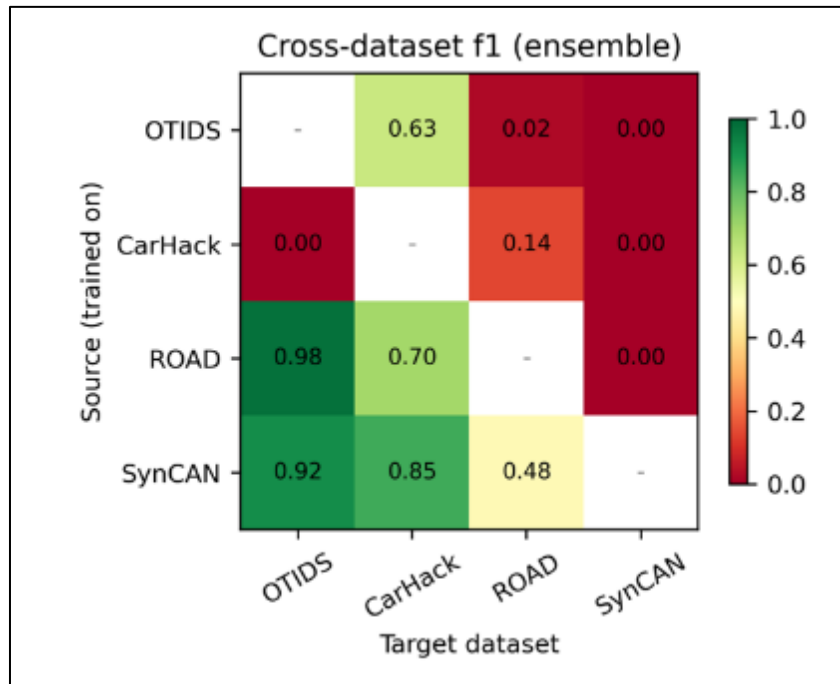


Figure 2: Cross-dataset F1, source (rows) to target (columns), ensemble model.

Table 3: Mean cross-dataset metrics over all 12 transfer pairs

Model	F1	Attack recall	Precision	Benign alarm rate
RF	0.345	0.379	0.593	0.364
Ensemble	0.394	0.376	0.618	0.215

5.3 The Representation Is the Bottleneck

By keeping the model fixed and only changing the representation, the mean transfer F1 for a fixed RF increases from 0.147 (for Rep-A and the identifier one-hot encoding) to 0.335 (for Rep-B and the identifier-agnostic encoding with local calibration). For each of the pairs, the underlying mechanism is visible: The import of identifier priors by Rep-A models, which are not relevant for a different vehicle, results in these models flagging or completely missing identifier groups. In contrast, the features of Rep-B are physical quantities (timing deviations, rate ratios, vocabulary novelty), whose semantics are preserved across vehicles. The cheapest way for transferable CAN IDS is not more complex models, but a good representation that does not memorize the vehicle.

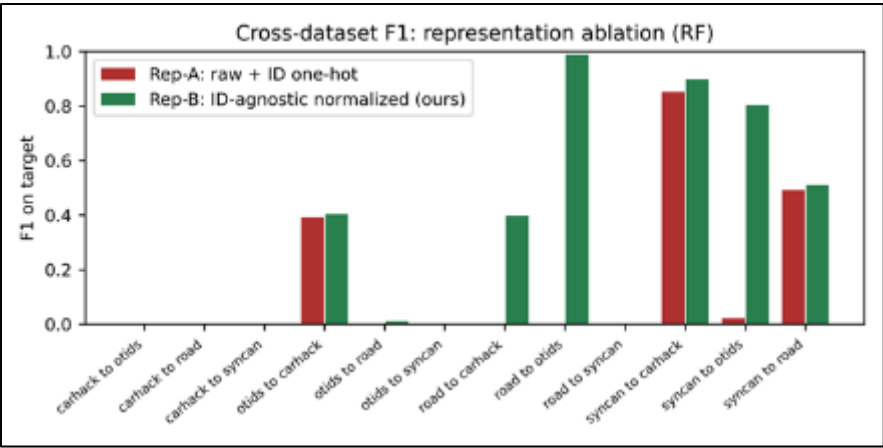


Figure 3: Cross-dataset F1, Rep-A versus Rep-B (RF).

Table 4: Representation ablation, per transfer pair (F1)

Transfer	Rep-A (ID one-hot)	Rep-B (ours)
CarHack to OTIDS	0.000	0.000
CarHack to ROAD	0.000	0.000
CarHack to SynCAN	0.000	0.000
OTIDS to CarHack	0.392	0.404
OTIDS to ROAD	0.000	0.012
OTIDS to SynCAN	0.000	0.000
ROAD to CarHack	0.000	0.400
ROAD to OTIDS	0.000	0.989
ROAD to SynCAN	0.000	0.000
SynCAN to CarHack	0.853	0.898
SynCAN to OTIDS	0.023	0.804
SynCAN to ROAD	0.492	0.511

5.4 Zero-Day Attack Families

In the scenario where an entire attack family is removed from the training data, the detection performance of the system is family-dependent (Figure 4). The performance for flooding-like attacks holds up using timing and rate features (mean held-out recall of DoS attacks is 0.32). On the other hand, families where the signature is based on identifier-specific payload substitution score close to zero. On average over all families, the held-out recall is 0.16. This is an honest number the field does not currently report. A detector that has been trained and certified on known attacks provides substantially less protection against structurally novel attacks.

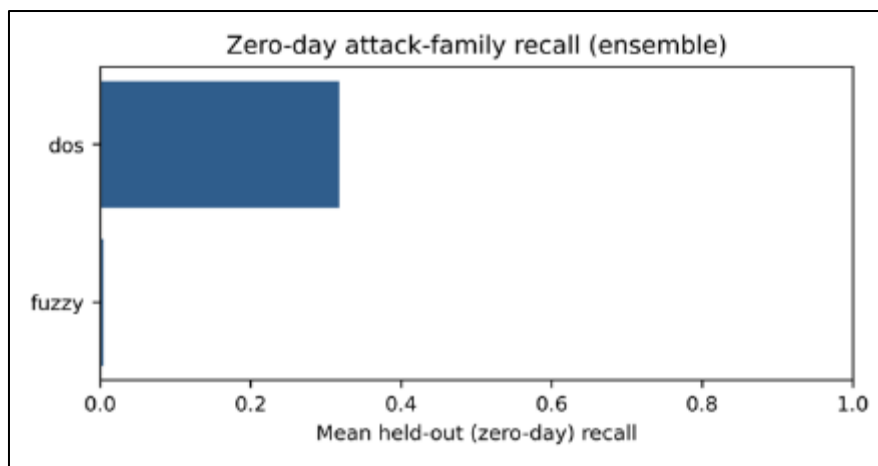


Figure 4: Held-out (zero-day) family recall, averaged over targets.

5.5 False Alarms in Operating Units

The number of benign alarms as expressed by the cross-dataset rates (Table 3, right column) translates into operator burden in terms of required tuning. The local calibration of Rep-B to keep this number under control as expressed by the thresholds, set on the basis of a vehicle's own statistics for benign cases, as opposed to those of the source vehicle, is key to this.

6 DISCUSSION AND LIMITATIONS

Our findings allow us to provide three recommendations for the community. First, reporting cross-dataset F1 scores in addition to in-dataset F1 scores should become mandatory, because a single in-dataset F1 says almost nothing about deployment. Second, identifier-agnostic, locally calibrated features should become the new default for the representation of automotive IDS data. These features are realistic to deploy and are based on data that is very easy to collect, namely minutes of unlabelled benign driving. Third, reporting of zero-day family recall as the main numbers in a paper, as opposed to including them in the supplementary material, should become the norm.

Limitations. The labeling of OTIDS' fuzzing behavior and ROAD's non-metadata messages uses documented heuristics for vocabulary novelty and injection identifier-based structures that, for rare events, might misclassify benign behavior; corpora differ in realism (SynCAN is fully synthetic); the model zoo is intentionally Classical (CPU-reproducibility); the single-model-to-ensemble escalation was applied once and it is possible that even larger gains could be obtained by future work on domain-adversarial training.

7 CONCLUSION

In this research, we evaluated four public CAN corpora. Single-dataset evaluation severely overstates what a CAN IDS can do. Models that achieve an in-dataset F1 score of 1.000 exhibit poor transferability, however, this research clearly showed that reducing identifier memorisation in favour of locally calibrated, identifier-agnostic statistics mitigates much of this detriment.

The average recall for zero-day families is 0.16. The primary innovation and notable contribution of this research is the clear demonstration that the representation, rather than the model's size, inhibits the success rate of CAN intrusion detectors in transferring among cars. Our identifier-neutral, locally adjusted representation corrects a significant portion of this limitation. It more than doubles the cross-dataset F1 score for the identical random forest without any labelled attack data from the target vehicle, and it enhances transfer pairs that had previously performed abysmally (ROAD to OTIDS: 0.000 to 0.989; SynCAN to OTIDS: 0.023 to 0.804). This indicates that a detector may be effectively modified for a new vehicle using only a few minutes of unlabelled benign traffic and CPU-scale training, rather than requiring a labelled attack campaign for each vehicle model (note that this represents a substantial advancement from this research to the AV Cybersecurity field). Since we assess zero-day recall for each attack family, the limitation of existing benchmarks transforms into a metric for evaluating future detectors. These outcomes and instruments provide the community with a method to construct CAN intrusion detectors that transfer more efficiently between vehicles, as well as a more rigorous and replicable criterion for their evaluation.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

- [1] Koscher K, Czeskis A, Roesner F, Patel S, Kohno T, Checkoway S, et al. Experimental security analysis of a modern automobile. In: 2010 IEEE Symposium on Security and Privacy; 2010; Oakland, CA, USA. IEEE; 2010. p. 447-62. doi:10.1109/SP.2010.34
- [2] Checkoway S, McCoy D, Kantor B, Anderson D, Shacham H, Savage S, et al. Comprehensive experimental analyses of automotive attack surfaces. In: Proceedings of the 20th USENIX Security Symposium; 2011; San Francisco, CA, USA. Berkeley (CA): USENIX Association; 2011.
- [3] Miller C, Valasek C. Remote exploitation of an unaltered passenger vehicle [Internet]. Black Hat USA; 2015 [cited 2026 Sep 27]. Available from: <https://illmatics.com/Remote%20Car%20Hacking.pdf>
- [4] Lee H, Jeong SH, Kim HK. OTIDS: a novel intrusion detection system for in-vehicle network by using remote frame. In: 2017 15th Annual Conference on Privacy, Security and Trust (PST); 2017; Calgary, AB, Canada. IEEE; 2017. doi:10.1109/PST.2017.00017. Dataset available from: <https://ocslab.hksecurity.net/Dataset/CAN-intrusion-dataset>
- [5] Seo E, Song HM, Kim HK. GIDS: GAN based intrusion detection system for in-vehicle network. In: 2018 16th Annual Conference on Privacy, Security and Trust (PST); 2018; Belfast, UK. IEEE; 2018. p. 1-6. doi:10.1109/PST.2018.8514157. Dataset available from: <https://ocslab.hksecurity.net/Datasets/car-hacking-dataset>
- [6] Koltai B, Ács G, Gazdag A. CAN we trust your results? A cross-dataset study of automotive IDS evaluation [Preprint]. arXiv:2606.30430; 2026 [cited 2026 Sep 27]. Available from: <https://arxiv.org/abs/2606.30430>
- [7] Verma ME, Bridges RA, Iannacone MD, Hollifield SC, Moriano P, Hespeler SC, et al. A comprehensive guide to CAN IDS data and introduction of the ROAD dataset. PLoS One. 2024;19(1):e0296879. doi:10.1371/journal.pone.0296879. Dataset available from: <https://zenodo.org/records/10462796>
- [8] Kidmose B, Kidmose A, Meng W. can-sleuth: sleuthing out the capabilities, limitations, and performance impacts of automotive intrusion detection datasets. Int J Inf Secur. 2025;24(5):193. doi:10.1007/s10207-025-01038-8
- [9] Song HM, Kim HR, Kim HK. Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network. In: 2016 International Conference on Information Networking (ICOIN); 2016; Kota Kinabalu, Malaysia. IEEE; 2016. p. 63-8. doi:10.1109/ICOIN.2016.7427089
- [10] Müter M, Asaj N. Entropy-based anomaly detection for in-vehicle networks. In: 2011 IEEE Intelligent Vehicles Symposium (IV); 2011; Baden-Baden, Germany. IEEE; 2011. p. 1110-5. doi:10.1109/IVS.2011.5940552
- [11] Taylor A, Japkowicz N, Leblanc S. Frequency-based anomaly detection for the automotive CAN bus. In: 2015 World Congress on Industrial Control Systems Security (WCICSS); 2015; London, UK. IEEE; 2015. p. 45-9. doi:10.1109/WCICSS.2015.7420322
- [12] Taylor A, Leblanc S, Japkowicz N. Anomaly detection in automobile control network data with long short-term memory networks. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2016; Montreal, QC, Canada. IEEE; 2016. p. 130-9. doi:10.1109/DSAA.2016.20
- [13] Lampe B, Meng W. can-train-and-test: a curated CAN dataset for automotive intrusion detection. Comput Secur. 2024;140:103777. doi:10.1016/j.cose.2024.103777
- [14] Rajapaksha S, Madzudzo G, Kalutarage H, Petrovski A, Al-Kadri MO. CAN-MIRGU: a comprehensive CAN bus attack dataset from moving vehicles for intrusion detection system evaluation. In: Symposium on Vehicle Security and Privacy (VehicleSec 2024); 2024; San Diego, CA, USA. Reston (VA): Internet Society; 2024. doi:10.14722/vehiclesec.2024.23043. Dataset available from: <https://archive.ics.uci.edu/dataset/1035/can-mirgu>
- [15] ETAS. SynCAN dataset (synthetic CAN bus data) [Internet]. GitHub; 2019 [cited 2026 Sep 27]. Available from: <https://github.com/etas/SynCAN>