Systematic Evaluation of Automotive Intrusion Detection Datasets

Arash Vahidi
arash.vahidi@ri.se
RISE Research Institutes of Sweden,
Digital Systems, Computer Science
Lund, Sweden

Thomas Rosenstatter*
thomas.rosenstatter@ri.se
RISE Research Institutes of Sweden,
Digital Systems, Mobility & Systems
Göteborg, Sweden

Nishat I Mowla
nishat.mowla@ri.se
RISE Research Institutes of Sweden,
Digital Systems, Mobility & Systems
Umeå, Sweden

ABSTRACT

Some current and next generation security solutions employ machine learning and related technologies. Due to the nature of these applications, correct use of machine learning can be critical. One area that is of particular interest in this regard is the use of appropriate data for training and evaluation. In this work, we investigate different characteristics of datasets for security applications and propose a number of qualitative and quantitative metrics which can be evaluated with limited domain knowledge. We illustrate the need for such metrics by analyzing a number of datasets for anomaly and intrusion detection in automotive systems, covering both internal vehicle network and vehicle-to-vehicle (V2V) communication. We demonstrate how the proposed metrics can be used to learn the strengths and weaknesses in these datasets.

CCS CONCEPTS

• **Security and privacy → Intrusion detection systems**.

KEYWORDS

automotive security, intrusion detection, data quality

ACM Reference Format:
Arash Vahidi, Thomas Rosenstatter, and Nishat I Mowla. 2022. Systematic Evaluation of Automotive Intrusion Detection Datasets. In *Computer Science in Cars Symposium (CSCS '22), December 8, 2022, Ingolstadt, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3568160.3570226

1 INTRODUCTION

A modern vehicle is a complex electrical system with a large number of intelligent nodes. These nodes communicate over multiple in-vehicle networks and through various interfaces to the outside world. For example, a vehicle may contain 100 to 200 electronic control units (ECUs), communicating across multiple network segments using Ethernet, FlexRay, CAN and wireless technologies, such as LTE, Bluetooth, Wi-Fi, and additional proprietary technologies. Figure 1 shows an example of an in-vehicle network (IVN) divided into the different network domains.

This growing complexity in conjunction with the introduction of connected and self-driving vehicles has put automotive security in focus. Strandberg *et al.* [27] provide a summary of 52 published attacks against vehicles, among which 37 are classified as high or critical attacks. Furthermore, taking cooperative driving and vehicle-to-everything (V2X) communication into account, multiple new interfaces must also be protected and secured.

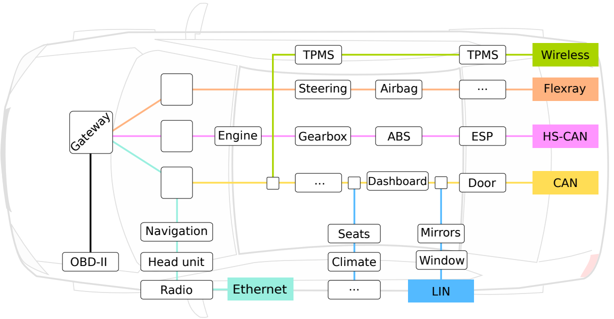


Figure 1: Example of internal network segments in a modern vehicle.

A significant problem with in-vehicle security is that some technologies were not designed with security in mind and offer minimal resistance once remotely or physically reachable to an attacker. For example, the CAN bus technology (see Figure 2) provides no security mechanisms and with a payload size of only 8 bytes adding one is next to impossible. New technologies (such as CAN-FD [46]) have been proposed to address shortcomings of CAN, however, due to technical or economical reasons it may not be possible to completely abandon CAN in near future.

Modern vehicles may also communicate with other vehicles (V2V communication), or to other entities (V2I communication) such as road-side units (RSUs). The communication between the vehicles themselves and to RSUs is established by forming a Vehicular Ad-Hoc Network (VANET). The properties of the communication between these entities bring different security challenges. For instance, security solutions for VANETs need to be able to cope with the ephemeral nature of VANETs where the vehicles are driving with different speeds and directions and thus join and leave the network constantly.

Automated monitoring systems, such as intrusion detection systems (IDS), have been proposed as a possible solution to the increasing threats and in future their presence may even be a regulatory requirement. Furthermore, the next generation of automotive IDSs may utilize artificial intelligence - and in particular machine learning - for anomaly detection. Such solutions often require properly curated datasets for training and evaluation. Past studies of IDS datasets for general network security have shown that minor flaws

[*] Corresponding author

CSCS '22, December 8, 2022, Ingolstadt, Germany

in datasets can greatly weaken a security solution [53]. Hence it is important that security researchers can analyze datasets with respect to properties relevant to security.

In this work, we investigate possible dataset requirements for security applications. We also analyze a number of commonly used automotive datasets, and explore whether those are suitable for use with automotive IDSs .

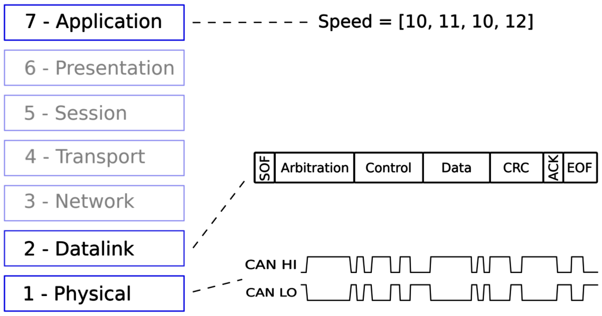


Figure 2: CAN OSI layers.

1.1 Contributions and outline

The main contributions of this work are as follows:

- We propose a systematic way to evaluate intrusion detection datasets for automotive systems.
- We evaluate a number of available automotive IDS datasets with respect to our proposed method.
- We highlight areas that must be improved in future datasets.

The remainder of this paper is organized as follows: Security in modern vehicles is discussed in Section 2, where we also discuss previous and related work. A set of IDS dataset requirements are proposed in Section 3, followed by analysis and discussion of deficiencies in available datasets in Section 4. Finally, we conclude our work with some observation and recommendations in Section 5.

2 AUTOMOTIVE SECURITY

Checkoway *et al.* [10] in 2011 and Miller and Valasek [41] in 2015 demonstrated how easily vehicles from that period could be compromised. More recent work by Yan *et al.* [61] and others [54] indicate that moderns vehicles are still insufficiently protected and security needs to be considerably improved. To ensure a better and more organized approach to security in road vehicles, ISO/SAE 21434 was introduced in 2021 to establish a common terminology and define a security process for all stages of a vehicle's life cycle. Furthermore, the UNECE Regulation No. 155 was created to provide a harmonized regulatory framework [1].

To develop a versatile security solution, a large array of attacks need to be considered in automotive security. These include, for instance, theft of the vehicle, causing physical harm to individuals, making unauthorized purchases using the vehicle's or the real owners' credentials, unauthorized tracking, reduction of vehicle performance, and the disruption of V2X communication. McCarthy *et al.* [37] and Karahasanovic *et al.* [29] discuss different attacks as well as the motivation driving such operations. Rosenstatter *et al.* [47] further propose a framework guiding developers in designing resilient systems capable of coping with this growing number of attacks and recover to the desired state.

2.1 Threat model

As noted earlier, the goals and the motivations of threat agents may vary greatly. Similarly, the incentives and the technical abilities of different attackers may vary significantly. While one could attempt to label the attackers (e.g. script kiddies, hackers, state actors) and use that to deduce their competence and resources, that approach may not be optimal for the analysis proposed in this work. We will therefore instead only consider the technical capabilities of the adversary against the vehicle, defined as the following:

- *Read capability*: Ability to observe the communication, e.g., eavesdropping.
- *Write capability*: Ability to inject malicious but otherwise valid messages.
- *Suppress capability*: Ability to suppress valid messages in the network.
- *Replace capability*: Ability to control another entity and its communication.
- *Direct access capability*: Ability to physically access the communication channel and operate it outside its specifications.

The reason for selecting these particular capabilities is to distinguish among the different ways attacks such as spoofing and denial of service can be implemented. Notice that for IVN communication, *write* and *replace* capabilities are similar to the idea of weakly and strongly compromised ECUs introduced by Cho and Shin [12]. The *direct access* capability assumes that the adversary has either compromised a specially privileged component (as demonstrated by Sagong *et al.* [48]) or has added their own hardware to the network. In the case of VANET, the wireless communication allows trivial *read* and *write* and relatively easy *suppress*, but rules out *direct access*.

Given the large number of attacks and attack variations, we also group similar attacks based on their type. This can be done according to the security attribute each attack violates, such as *Confidentiality, Integrity* and *Availability* (CIA) [44]. For this work, we chose to group attacks according to the Microsoft STRIDE threat categories [40], which are well established in the automotive domain [26, 36, 51].

- *Spoofing*: Impersonation of something/someone, for example sending manipulated brake signals.
- *Tampering*: Modification of data or code, for example manipulating ECU firmware.
- *Repudiation*: Deniability of actions, for example deleting or overwriting sensitive data to hinder or complicate forensics.
- *Information disclosure*: Unauthorized access to information, for example extraction of sensitive vehicle data.
- *Denial of Service*: Refuse or reduce access to a service, for example reduce vehicle performance by generating large amounts of bogus traffic.
- *Elevation of Privilege*: Get unauthorized access, for example by using an implementation flaw in the infotainment system.

For example, flooding and blackhole attacks are both different implementations of *Denial of Service*.

2.2 Intrusion detection systems

Automotive systems have a number of inherent security problems that are challenging to overcome. Strandberg *et al.* [27] list a number of security concepts and patterns for improving security and resilience in automotive systems. Alongside standard security practices, such as encryption and access control, the authors also propose monitoring systems for detecting unwanted behavior. In future such security solutions may also be required by law. The UNECE Regulation No. 155 on automotive cybersecurity states that vehicle manufacturers shall implement measures to *"detect and prevent cyber-attacks"*, and shall support monitoring *"with regards to detecting threats, vulnerabilities and cyber-attacks"* [1].

To detect security threats, an IDS monitors events to detect unwanted behavior in real-time or off-line. Two commonly used types are host-based IDS (HIDS) which mainly monitor internal system events (*e.g.*, file access and system calls), and network-based (NIDS) that monitor network traffic. The focus of this work is NIDSs, although the discussion may also apply to HIDSs.

The two main IDS implementation paths are *signature-based* that utilizes a knowledge database describing wanted/unwanted behavior, and *anomaly-based* that observes deviations from a learned normal behavior [4]. The latter has gained popularity due to the advances in machine learning. Advantages of this approach include reduced need for expert knowledge (to construct a signature database) and a theoretical possibility to detect new attacks.

A possible disadvantage of anomaly-based IDSs is the need for high quality data used for training and evaluation. Previous research has demonstrated that unless this data meets certain quality requirements, the produced IDS will struggle to detect many real-world security threats [39, 53]. This subject has been discussed in length for general NIDSs, but until now has been somewhat neglected in the automotive world. Our work aims to address this issue by highlighting relevant dataset issues.

2.3 Datasets

The following terminology is used throughout this paper. A *dataset* D is normally a collection of one or more tables where each row is a data entry and columns represent variables. Consider for example the Iris flower dataset [17], shown in Table 1.

This dataset may be used to find a suitable model $\theta : X \longrightarrow Y$, to predict flower type Y based on the input measurements X. More formally, $x = [x_1, \cdots, x_n] \in X = \mathbb{R}^n$ is a vector of n *features* (*e.g.*, sepal length) the set of which we denote $Features(D)$. In a classification problem such as this, $y \in Y = \{c_1, \cdots, c_m\}$ is a discrete *label* (e.g. Iris setosa). Finally, we let D_{c_i} to represent the subset of D with the label c_i and $|D|$ to be the size of D.

A security dataset is usually a binary classification problem, which in this work has the two labels A (attack) and B (benign). Thus D_A and D_B would represent malicious and non-malicious entries in the dataset, $|D_A|$ and $|D_B|$ would denote their respective sizes, and so on.

For use in machine learning, a dataset should meet a number of requirements. For example, the selected features must be relevant to the task at hand and the number of different classes in the dataset should be approximately equal. We will revisit this in Section 3.

Table 1: Extract from the Iris flower dataset.

X				Y
Sepal length	Sepal width	Petal length	Petal width	Specie
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
6.2	3.4	5.4	2.3	Iris virginica
6.0	2.9	4.5	1.5	Iris versicolor
		. . .		

The Iris flower dataset and many similar popular datasets model static problems that do not change over time. Unfortunately, this assumption is almost never true for security applications where an intelligent attacker is able to improve and adapt. Hence, security datasets must meet a different set of requirements that are sometimes contradictory to the ones used for normal datasets. We will discuss these requirements in detail in the following sections.

2.4 Related work

The *KDD CUP 99* dataset [22] is a pioneering network dataset published by the MIT Lincoln Labs as part of the DARPA Intrusion Detection Evaluation Program. This dataset was extracted from multiple weeks of traffic in a network environment that supposedly resembled a US Air Force LAN. While widely used in IDS research, McHugh [39] noticed a number of shortcomings in this dataset, including poor documentation and unrealistic network traffic. Tavallaee *et al.* [53] noted that this dataset had not resulted in significant improvement in real-world IDS solutions, which the authors after careful statistical analysis attributed to multiple issues with the data. These included disparities between captured and expected LAN traffic, missing packets, redundant entries, and attack data imbalance. Kenyon *et al.* [30] revisited this subject by examining this and 26 other IDS datasets two decades after its publication. The authors conclude that many are either flawed, not fit for the desired purpose or simply outdated.

To better generate synthetic attack data, Cordero *et al.* [13] investigated common defects in IDS datasets, which included the use of impossible artificial anomalies, predictable patterns, and unrealistic cleanness.

Quantitative analysis of datasets can be useful to detect certain problems. Ho and Basu [23] proposed a number of computationally cheap metrics for measuring dataset complexity. Xu *et al.* [60] investigated the idea of usable information, $\mathcal{V}$-information, as an extension of Shannon's information theory. The authors proposed a mechanism for efficiently predicting this quantity in a machine learning context. Ethayarajh *et al.* [15] used this to estimate the difficulty of some natural language processing datasets which would allow researchers to compare datasets and models and to study feature contribution.

Multiple attempts have been made to improve CAN security without significantly altering the protocol. For example, Nilsson *et al.* [43] suggest using a 64-bit MAC for authentication and integrity control. The proposed solution computes a MAC for every 4 frames, which replaces the CRC for the next 4 frames. Groll and Ruland [19] consider encrypted CAN communication, where all ECUs in a security domain share a symmetric key, with initial key distribution

Table 2: Data readiness with weighted deficiencies [8].

Band	Weight	Deficiency
C	40	Parseability
	25	Data storage
	15	Decoding
	10	Data formats
	10	Disjoint datasets
B	20	Column types
	30	Missing values
	20	Inconsistent data entries
	10	Duplicated records
	20	Meaningful values
A	20	Interpretable values
	20	Feature scaling
	20	Outlier detection
	30	Feature selection
	10	Coverage gap
AA	40	Legal violations
	40	Security risks
	20	Bias detection
AAA	-	None

handled by a dedicated ECU. Khodari *et al.* [31] proposed verifying ECU correctness by using attestation in a distributed fashion, where an ECU may appraise multiple other ECUs, and vice versa.

Verma *et al.* [58] investigate a number of automotive IDS datasets and note a number of problems including incorrect documentation or labels and very noisy attacks that can be identified with very simple methods. Swessi and Idoudi [52] provide a recent survey of automotive security datasets. They evaluate each dataset based on 11 criteria and conclude with recommended datasets for in-vehicle and inter-vehicle machine learning-based IDSs.

3 IDS DATASET REQUIREMENTS

The quality of the training dataset directly affects real-world IDS performance, making dataset quality assurance a crucial requirement. This can become a complex and time consuming task, but utilizing the concept of *data readiness* we are able to provide a simple and systematic method to assess datasets.

Lawrence proposed the idea of *data readiness levels* [34] by placing data within one of three main bands: (C) dataset exists, (B) data exists and is a faithful representation, and (A) correct data exists and is appropriate for this task. To allow a generic quality assessment without the need for domain experts, Castelijns *et al.* [8] proposed the alternative definition presented in Table 2. Under this new definitions the bands are constructed to allow analysis at different stages of the project: (C) data is not yet read, (B) data is not yet analyzed but minor errors have been addressed, (A) data is ready for deeper analysis, (AA) use of data is considered. Based on the weighted *deficiencies* shown in Table 2, each band is given a score between 0 and 1. The appropriate band is then selected after choosing a threshold (the authors suggested 0.85).

Table 3: Proposed data readiness for IDS applications.

Band	Weight	Deficiency
C	30	Dataset documentation
	30	Objective
	20	Parseability
	20	Dataset age
B	40	Format correctness and consistency
	20	Dataset size
	20	Completeness
	20	Label inclusion and correctness
A	20	Class balance
	30	Attack documentation
	20	Security coverage
	30	Attack realism
AA	40	Dataset realism and diversity
	20	Feature context and documentation
	20	Difficulty
	20	Transformation and anonymization
AAA	-	None

3.1 IDS data readiness

Unfortunately, the deficiencies proposed in Table 2 are not optimal for intrusion detection and possibly many other security applications. For instance, removing outliers may remove the very anomalies the IDS is meant to identify.

Fortunately, it is possible to use the dataset readiness concept with other deficiencies and weights. Hence, we propose the alternative deficiencies shown in Table 3 for use with IDS datasets. We chose the deficiencies and weights to highlight areas we found to be problematic. In the remainder of this chapter we will discuss the motivation behind this selection. Furthermore, in Chapter 4 we will demonstrate how to quantify these deficiencies when examining IDS datasets.

3.2 Dataset documentation

We consider documentation to be one of the most important parts of a security dataset, and therefore propose three documentation deficiencies:

(1) general documentation describing content and origin,
(2) description of the attacks in the dataset and how they were executed/recorded, and
(3) description of the features (*e.g.*, origin, meaning, range) and their physical context (*e.g.*, how vehicle speed, engine speed and gear are related).

Motivation: Proper documentation is needed to assess if its content and creation methodology aligns with our requirements. Furthermore, information about the included attacks is needed to judge its contribution to an IDS. Finally, information about features and their relations to each other and their surrounding may be needed to allow explainability, or to create a context-aware IDS [59].

3.3 Labels

Each entry in an IDS dataset may be given a *label* for identifying whether that entry is benign or an attack: $Y = \{B, A\}$. This information is usually needed for IDS training and should be included in the dataset.

Motivation: While this information is often crucial for training, many datasets lack labels or provide incomplete or incorrect labels.

3.4 Parseability, correctness and consistency

We simplify the numerous deficiencies in the original readiness model to only three:

(1) data should be stored in an appropriate machine/human-readable format (*e.g.*, PCAP or CSV rather than SQL databases),
(2) all entries should be correctly formatted (*e.g.*, no corrupt entries), and
(3) use a single data format for all entries.

Motivation: In the original data readiness model many deficiencies were related to data representation and validity. The assumption was that a good dataset should be in a format that can be used by non-domain experts with minimal effort. Hence, the dataset authors should handle tasks such as feature selection and data normalization and provide the data in a format ready for immediate use in a machine learning framework.

However, while studying public IDS implementations we observed significant disagreement in feature selection and representation. Hence, we believe dataset authors should rather focus on providing correct, complete, and unaltered data. Extracting and transforming data to the desired format is then done by the IDS developers. This should also make the dataset more future-proof as more IDS designs and attack variations are discovered.

3.5 Age, size and objective

To ensure that the included data is large enough and somewhat up-to-date we propose the size and age deficiencies. Another important issue to consider is whether the dataset objective is compatible with ours (*e.g.*, IDS development).

Motivation: The threat landscape can change very quickly, and a dataset with obsolete information is not very useful. While not perfect, dataset age might be a useful indicator for this. It is also important that the dataset contains enough data for training and evaluation. An often used rule of thumb is that dataset size should be at least an order of magnitude larger than the number of parameters in the model.

Finally, it is important that the objective of the dataset *aligns* with (but not necessarily *identical* to) ours. For example, the HCRL Driving dataset (see Section 4) contains a large amount of benign data and no cybersecurity issues. However, this may be quite useful for validating IDS robustness and reducing false-positives under different driving styles and road conditions.

3.6 Completeness, transformation and anonymization

As already noted in Section 3.4, a security dataset should be *complete* in the sense that no key features or entries have been discarded. Similarly, data should not be irreversibly *transformed* (*e.g.*, changing timestamp granularity from milliseconds to seconds). This also includes attempts to *anonymize* the data, although we understand that this may be in conflict with privacy requirements.

Motivation: In the KDD99 dataset background traffic that was initially considered inconsequential turned out to be important to real-world IDS performance [39]. In addition, if an IDS observes only a subset of the system events, an attacker may be able to craft an attack invisible to the IDS and thereby avoid detection.

3.7 Dataset and attack realism

Dataset realism and diversity considers whether a dataset is a good representation of the problem at hand. For example, a network dataset may include unrealistic synthetic traffic, or ignore some network protocols, or only capture data from some network nodes. The included security attacks can also exhibit similar problems, thus we consider *attack realism* as a separate deficiency. Furthermore, attack diversity is examined as part of *attack coverage*.

Motivation: Kenyon *et al.* [30] recognize three ways to create intrusion datasets: capturing live data from a real-world event, generating synthetic data in a simulated environment, and a hybrid approach with both real and synthetic data. The last two are more likely to contain unrealistic data. To complicate matters, sometimes benign and attack data use different methodologies. For example, it is not uncommon to record live benign data from a vehicle and later insert synthetic attacks into the dataset. Hence we consider general dataset realism and attack realism as two separate deficiencies.

3.8 Security coverage

Since an IDS that only detects one type of attack is seldom useful, a comprehensive dataset (or collection of datasets) should include a wide variety of attacks. The *security coverage* deficiency captures this by evaluating attack classes and implementations in the dataset.

Sharafaldin *et al.* [50] propose the metric *attack diversity*, which is the number of attacks included in the dataset out of a catalog of seven groups of attacks. As that catalog covers general computer security issues (*e.g.*, browser attacks), we instead propose *attack coverage* as the number of different security threats covered in a dataset. While not optimal, we use STRIDE for this task:

$$\text{attack coverage} = \frac{|A \cap C|}{|C|} \ ,$$

where A is the set of threats covered in the dataset and $C \subseteq \{S, T, R, I, D, E\}$ is the catalog of threats the IDS is meant to detect.

Note that a threat class such as spoofing may include a wide variety of attacks, but it is not practical to cover every single attack implementation in a dataset (specially as many are not even known to us). Hence. we instead consider what the attacker theoretically could do by considering the attacker capabilities covered in the dataset:

$$\text{capability coverage} = \frac{|AC \cap CC|}{|CC|} \ ,$$

where AC is the set of included capabilities and $CC \subseteq \{Read, Write, \cdots\}$ is the capability catalog. To capture both, we define the *security coverage* deficiency as

$$\text{security coverage} = \frac{\text{attack coverage} + \text{capability coverage}}{2}$$

Motivation: Many datasets consider a very specific type of attack or a very specific implementation of an attack, resulting in poor real-world IDS performance. By understanding this deficiency we are better positioned to understand how the IDS may perform against attacks seen in the wild.

3.9 Dataset difficulty

The *difficulty* deficiency considers the hardness of the problem the dataset captures. However, as this is not a trivial task, a simple heuristic to approximate dataset *complexity* may be used instead. For example, Ethayarajh *et al.* [15] evaluate difficulty of natural language datasets under a given classifier by measuring a dataset's *useful information* [60]. For IDS datasets a simple heuristic such as the *volume of overlapping regions* [23] (for calculating the intersecting volume of *attack* and *benign* entries) may be sufficient to detect some problems:

$$\prod_{f \in Features(D)} \frac{MIN(max(f_A), max(f_B) - MAX(min(f_A), min(f_B))}{MAX(max(f_A), max(f_B) - MIN(min(f_A), min(f_B))}$$

Motivation: McCoy *et al.* [38] noted that by only learning shallow heuristics from training data, machine learning models could perform well on training data but fail in real-world. Similarly, Verma *et al.* [58] observed this issue in many automotive IDS datasets due to the clumsy way some attacks were implemented. Many such issues may be detected when examining dataset difficulty.

3.10 Data balance

In machine learning it is normally preferred that all classes are of roughly equal size [9]. Unfortunately, security datasets often contain significantly more benign than malicious entries (*i.e.*, $|D_B| \gg |D_A|$). To highlight this problem we use the *data balance* deficiency:

$$\text{data balance} = \sqrt{1 - \left| 2\frac{|D_A|}{|D|} - 1 \right|},$$

where a 0/100 and a 50/50 split yield 0 and 1 respectively.

Motivation: While there exist techniques to somewhat improve training and evaluation of imbalanced datasets (*e.g.*, SMOTE [16]), the importance of data balance is still something dataset and IDS authors should consider. The real-world implication of dataset imbalance in a security dataset is not always easy to measure and may depend on other subjects [1]. Therefore, we define a somewhat forgiving method for quantifying this deficiency.

4 ANALYSIS OF AUTOMOTIVE IDS DATASETS

In the following we will analyze a number of available datasets in light of the proposed data readiness deficiencies. To ensure a fair assessment, we first formulated a simple method to quantify each deficiency, which is presented in Table 4. We then split the datasets into three overlapping sets such that each dataset was analyzed by two out of the three authors. Note that Table 4 is only provided as an example and is based on what the authors judged as relevant information for their particular application.

4.1 Examined datasets

For analysis we consider the datasets presented in Table 5, which covers both in-vehicle and VANET/V2V. Most in-vehicle datasets focus on CAN traffic, and are often recorded from the vehicle diagnostics port (OBD-II). The SIMPLE dataset [18] differs in this regard as it contains bus voltage level measurements from an internal CAN bus. VANET datasets originate mostly from simulations, as attacks are more complex and may require several VANET enabled vehicles. One exception is VDoS-LRS [45]. Many VANET simulations focus on misbehavior, meaning that malicious vehicles provide incorrect information or spoof the presence of other vehicles with the aim to mislead the surrounding vehicles. Authors may provide simulation configuration and setup, which has the advantage that datasets can be extended or adapted to one's specific needs and requirements.

In Table 5, we also included datasets not primarily created for intrusion detection that may still be valuable in security contexts, namely (i) the *Halmstad GCDC* dataset [3] contains internal events and external communication of a real VANET environment, and (ii) the *HCRL Driving Dataset* [32] contains approximately 23 hours of driving one vehicle with 10 different drivers. Finally, the Next Generation Simulation (NGSIM) dataset [56] does not contain any attacks, but is often manipulated for use of misbehavior detection.

It should be noted that this list does not include datasets for autonomous driving, e.g., DriveTruth [42] and nuScenes [7]. The later also provides an overview of other automotive driving datasets.

4.2 Dataset documentation

The three documentation deficiencies in IVN and VANET datasets are shown in Figure 4. Most data authors provided information about the dataset and the attacks, however, we observed that very few described both, selected features and their context. Furthermore, the provided documentation was sometimes ambiguous, forcing us to make some assumptions (which may have affected the interpretation of the other deficiencies).

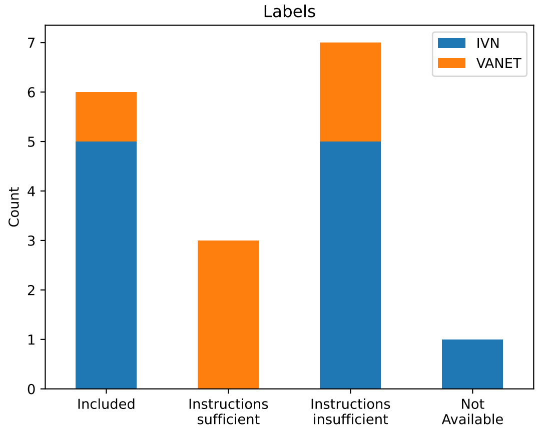


Figure 3: Deficiency for labels.

[1] For example, some datasets include a large benign component provided for a different purpose. Dataset balance can also vary greatly depending on data representation. For example, using each entry as a data point has most likely a different balance than using a window of N entries.

Table 4: Proposed data readiness for IDS applications with evaluation metrics provided as an example.

Band	Deficiency	Metric used in the evaluation
C	Dataset documentation	+0.5 if files described +0.5 if environment and context described
	Objective	+0.5 if objectives explained +0.5 if objective aligns with ours
	Parseability	1.00 if standard format (e.g., CSV, JSON, PCAP) 0.75 if non-standard but readable format 0.50 if data requires pre-processing 0.00 otherwise
	Dataset age	$1 + (\text{year of publication} - \text{start})/N$ (limited to $[0, 1]$, $N = 5$, $start = 2021$)
B	Format correctness and consistency	+0.5 if no errors +0.5 if single format
	Dataset size	1.00 if > 1,000,000 samples/attack type 0.75 if > 100,000 samples/attack type 0.50 if > 10,000 samples/attack type 0.25 if > 1,000 samples/attack type 0.00 otherwise
	Completeness	**IVN (CAN):** 1.00 if entire CAN frame included 0.75 if only few non-essential fields are removed 0.50 if only ID + DLC + CAN data available 0.25 if ID + padded payload 0.00 otherwise. **VANET:** 1.00 if all data included 0.50 if some background data removed 0.00 otherwise.
	Label inclusion and correctness	1.00 if labels included 0.50 if instructions provided 0.25 if incorrect or incomplete instructions provided 0.00 otherwise.
A	Class balance	See Section 3.10
	Attack documentation	+0.5 if attacks described +0.5 if attack setup described
	Security coverage	See Section 3.8
	Attack realism	1.00 if real attacks in the wild 0.75 if real attacks by authors 0.50 if realistic synthetic attacks 0.00 otherwise
AA	Dataset realism and diversity	+0.50 if multiple different vehicles +0.25 if multiple driving situations +0.25 if multiple drivers
	Feature context and documentation	+0.25 if features explained +0.75 if context explained
	Difficulty	See Section 3.9
	Transformation and anonymization	+0.50 if not anonymized +0.50 if not modified at all
AAA	None	–

4.3 Labels and class balance

As evident from Figure 3, some datasets did not provide labels or instructions to correctly reconstruct the labels. Unfortunately, lack of labels disqualifies a dataset from most applications. In fact, we could not fully complete our investigations for multiple datasets as some deficiencies could not be evaluated without access to labels.

For class balance, we also included datasets with partial or possibly incorrect labels to obtain more samples. The results can be seen in Figure 5, where IVN datasets seem to perform somewhat worse. The reason we identified is that most VANET datasets are based on simulations which allow for easier labeling. Note, however, that except in extreme cases datasets with low balance score may still be perfectly usable, although it is important to address the imbalance during training and evaluation.

4.4 Parseability, correctness and consistency

While such issues seem to have been dominant in earlier studies (as evident from Table 2), we did not note any such issues among the examined datasets. In fact, the biggest problem we encountered was that a small number of datasets employed two different (yet still very similar) file formats.

4.5 Age, size and objective

As shown in Table 4, the age deficiency decreases linearly towards zero after the *start* year. Figure 5 indicates that VANET datasets

Table 5: Overview of the analyzed intrusion detection datasets. Notes: incomplete dataset (*), the simulation is provided as description (desc), with source code (sc) and/or with dataset (da).

Dataset	Data Type		Source	Objective	Year	Ref.
	Benign	Attack				
In-Vehicle datasets						
HCRL OTIDS	real	real	CAN/OBD-II	Intrusion D.	2017	[35]
HCRL Car-Hacking Dataset	real	real	CAN/OBD-II	Intrusion D.	2018	[49]
HCRL Survival	real	real	CAN/OBD-II	Intrusion D.	2018	[20]
TU Eindhoven v2	real	synthetic	CAN/OBD-II	Intrusion D.	2019	[14]
SIMPLE	real	real	CAN voltage	Intrusion D.	2019	[18]
SynCAN	synthetic	synthetic	CAN	Intrusion D.	2019	[21]
ORNL	real	real/synthetic	CAN/OBD-II	Intrusion D.	2020	[58]
CrySyS	real	synthetic	CAN + GPS	Intrusion D.	2021	[11]
Hisingen	real	real	CAN/OBD-II	Intrusion D.	2021	[24]
Bi2022*	real	real	CAN/OBD-II	Intrusion D.	2022	[6]
VANET datasets						
Belenko2018	synthetic	synthetic	Sim(desc)	Intrusion D.	2018	[5]
VeReMi	synthetic	synthetic	Sim(da,sc,desc)	Misbehavior D.	2018	[57]
Lastinec2019	synthetic	synthetic	Sim(desc)	Misbehavior D.	2019	[33]
VeReMi Extension	synthetic	synthetic	Sim(da,desc)	Misbehavior D.	2020	[28]
VDoS-LRS	real	real	IEEE 802.11g	Intrusion D.	2020	[45]
VDDD	synthetic	synthetic	Sim(desc)	Intrusion D.	2021	[2]
Iqbal2021	synthetic	synthetic	Sim(da,desc)	Misbehavior D.	2021	[25]
Other datasets						
HCRL Driving Dataset	real	–	CAN/OBD-II	Driving behavior D.	2016	[32]
NGSIM	real	–	Vehicle trajectories/videos	Modelling traffic	2016	[55]
Halmstad GCDC Data	real	–	V2V comm.	Misbehavior D.	2018	[3]

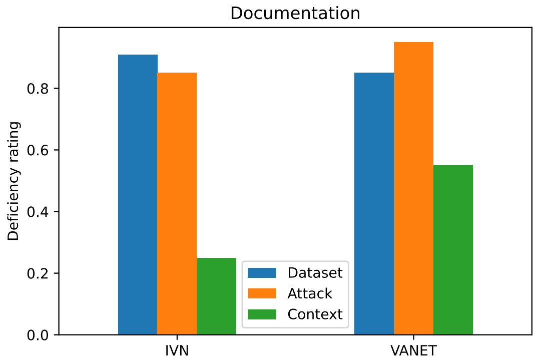


Figure 4: Deficiency for documentation.

are slightly more recent, which seems to agree with the current research activities in the automotive domain.

For the size deficiency a model with 10^5 parameters was assumed, requiring at least 10^6 data samples to train. Furthermore, the size was calculated separately for each type of attack in the dataset and the smallest number was used to calculate the size deficiency. It was noted that while IVN datasets were often smaller than ideal, the VANET datasets were often sufficiently large.

The objective deficiency was significantly less exciting as almost all evaluated datasets were handpicked for their use in IDS applications, and therefore received a perfect score of 1.0.

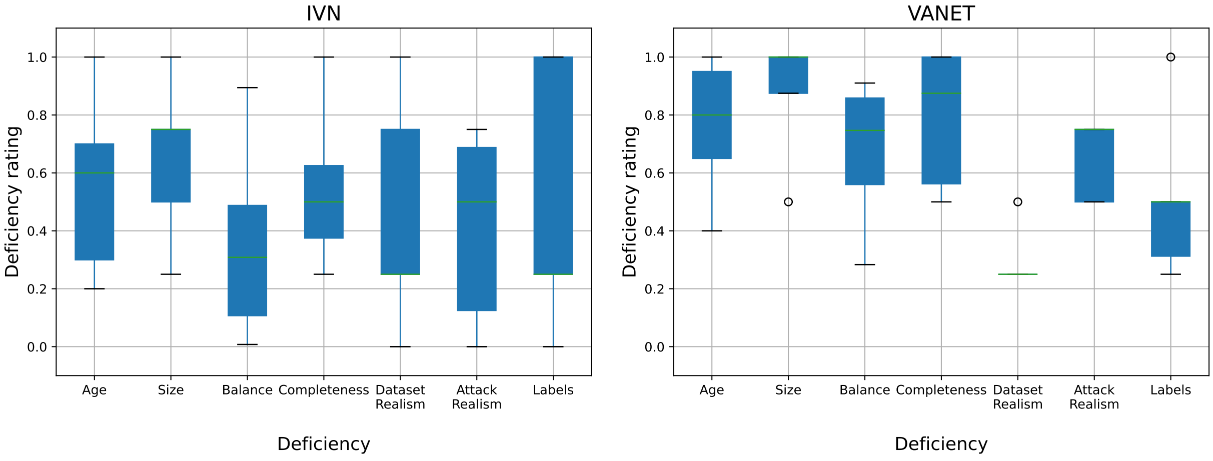
4.6 Completeness, transformation and anonymization

Very few datasets were complete according to the criteria described in Table 4. For example, VANET datasets sometimes had background traffic or essential features removed. IVN datasets often excluded important fields such as CRC, control flags and length, sometimes only providing application layer information (see Figure 2). As such, IVN datasets often lacked the information crucial for detecting threat agents with *replace* and *direct access* capabilities.

We also investigated whether any datasets had been anonymized or otherwise modified, including attempts to normalize the data. Unfortunately, this was not uncommon in IVN datasets while VANET dataset were often provided in their raw original format, most likely thanks to originating from simulations.

4.7 Realism and diversity

To quantify dataset realism and diversity, we investigated whether datasets include data from multiple vehicles, drivers and driving situations. As seen in Figure 5, few IVN datasets and even fewer VANET datasets did this.



(a) **(b)**

Figure 5: Various deficiencies for IVN and VANET datasets.

Attack realism was analyzed according to Table 4. IVN datasets were shown to contain more unrealistic synthetic attacks despite VANET datasets being mostly simulations. To the best of our knowledge no dataset contained real attacks recorded in the wild.

4.8 Dataset difficulty

To better understand the need for difficulty, we first constructed a simple experiment. Four different models (including ResNet-IDS proposed by Seo *et al.* [49]) were trained against the HCRL Car-Hacking DoS dataset. After training, all malicious entries were modified to use a different high-priority sender ID (*e.g.*, 2 instead of 0). As shown in Table 6, this minor change (a single input bit was changed without otherwise affecting the attack) resulted in a significant IDS performance degradation. This demonstrates existence of shallow heuristics in the dataset (see Section 3.9) that could result in a very fragile IDS, vulnerable to trivial evasion attacks.

The estimated difficulty based on dataset complexity is shown in Figure 6. This simple method was able to capture some dataset issues, including the aforementioned problem.

4.9 Security coverage

We observed that IVN datasets mainly cover denial of service (DoS) and spoofing attacks aiming to affect vehicle functions like speed. Similarly, VANET datasets focus on changing the behavior of surrounding vehicles by performing routing or spoofing attacks.

Table 6: Performance degradation due to flawed DoS training dataset.

Model	Parameters	Training time [s]	F1 before	F1 after
MLP	8.5×10^4	9	0.99760	0.79480
CNN 1D	1.5×10^3	22	0.99710	0.01343
CNN 2D	1.0×10^3	400	0.99240	0.00021
ResNet	1.7×10^6	3000	0.99920	0.20974

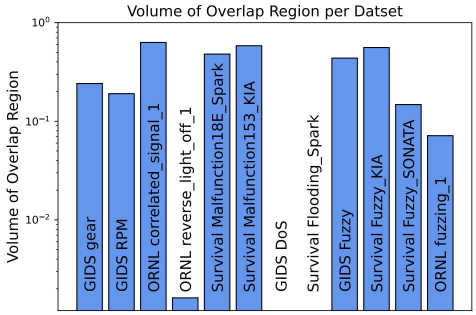
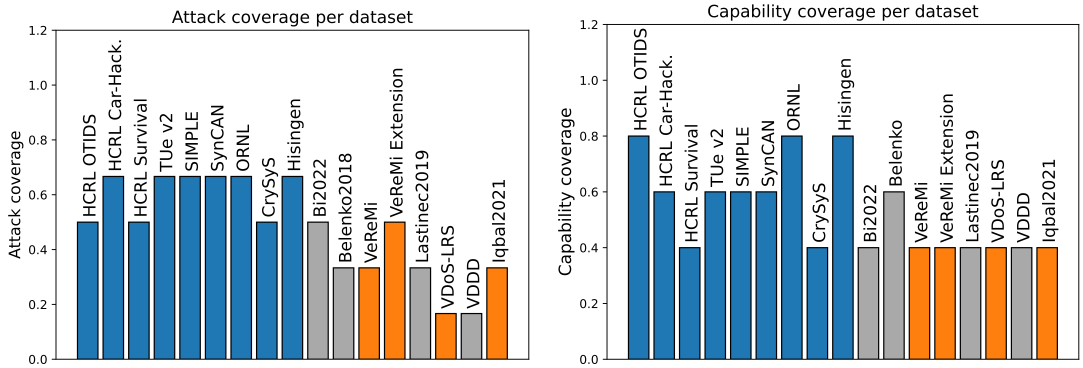


Figure 6: Volume overlap for some IVN datasets.

Attack and capability coverage for all datasets is shown in Table 7. Spoofing and DoS are most common while repudiation attacks are not included in any dataset and information disclosure was included in one dataset (TUe v2) where they exploited diagnostic functions.

Figure 7 (a) shows the calculated attack coverage for each dataset, with the highest score being 0.67. This analysis includes all six STRIDE threats, although one could consider excluding repudiation from a NIDS attack catalog. IVN datasets cover more types of threats than VANET datasets, which mainly consider misinformation and DoS. Two VANET datasets (VDDD and VDoS-LRS) focus exclusively on variations of DoS. Such detailed datasets are also important for IDS training, however, the aim of this assessment is to evaluate how comprehensive datasets are.

The capability coverage shown in Figure 7 (b) was calculated according to Section 3.8. All datasets consider *read* and *write* capabilities while *suppress* and *replace* are much less common. To the best of our knowledge no datasets covered *direct access* capabilities, although in the case of VANET one could argue this capability is not applicable due to wireless communication. Furthermore, if we only consider control of other entities such as other vehicles and RSUs (thus excluding intentional malicious modifications to the



(a)

(b)

Figure 7: Attack coverage (a) according to STRIDE and (b) the capability coverage rating. The coverage of datasets and simulations in gray is only based on their description.

Table 7: Security coverage of intrusion detection datasets.

Dataset	Threat						Capability				
	S	T	R	I	D	E	Re	Wr	Sup	Rep	Dir
In-Vehicle datasets											
HCRL OTIDS	S	–	–	–	D	E	Re	Wr	Sup	Rep	–
HCRL Car-Hacking Dataset	S	T	–	–	D	E	Re	Wr	Sup	–	–
HCRL Survival	S	–	–	–	D	E	Re	Wr	–	–	–
TU Eindhoven v2	S	–	–	I	D	E	Re	Wr	Sup	–	–
SIMPLE	S	T	–	–	D	E	Re	Wr	Sup	–	–
SynCAN	S	T	–	–	D	E	Re	Wr	Sup	–	–
ORNL	S	T	–	–	D	E	Re	Wr	Sup	Rep	–
CrySyS	S	T	–	–	D	–	Re	Wr	–	–	–
Hisingen	S	T	–	–	D	E	Re	Wr	Sup	Rep	–
Bi2022	S	–	–	–	D	E	Re	Wr	–	–	–
VANET datasets											
Belenko2018	S	–	–	–	D	–	Re	Wr	Sup	–	–
VeReMi	S	T	–	–	–	–	Re	Wr	–	–	–
VeReMi Extension	S	T	–	–	D	–	Re	Wr	–	–	–
Lastinec2019	S	T	–	–	–	–	Re	Wr	–	–	–
VDoS-LRS	–	–	–	–	D	–	Re	Wr	–	–	–
VDDD	–	–	–	–	D	–	Re	Wr	–	–	–
Iqbal2021	S	T	–	–	–	–	Re	Wr	–	–	–

attacker's own vehicle), no VANET dataset considers the *replace* capability. In conclusion, the available IVN and VANET datasets cover only a fraction of what an attacker is able to do.

It must be pointed out that our security coverage classification was based on the available documentation, which was sometimes unclear or ambiguous and left room for interpretation. As such, it may contain some errors.

5 CONCLUSIONS

Use of machine learning in security requires careful analysis of various aspects of the process. As noted in Section 2.4, the issue of dataset quality has been extensively studied in the area of IT network intrusion detection. In this work, we provide a similar analysis for automotive systems with a number of in-vehicle network and VANET datasets. More importantly, we propose a data readiness variant for systematic analysis of security datasets. This uncovers a number quality issues, such as:

1. While general documentation is usually available, information about context, attack type and attacker capabilities is often missing or ambiguous.
2. Labels are sometimes missing, incomplete or incorrect, which disqualifies the dataset from many IDS applications.
3. Many times the included attacks are not very realistic and to the best of our knowledge no dataset includes real attacks recorded in the wild.
4. IVN datasets are often limited to one type of network (CAN) recorded from a diagnostic port that may not fully represent the network (*e.g.*, due to internal firewalls and proxies).

Using the recommended data readiness threshold of 0.85, only one IVN dataset and two VANET datasets reached band A, and none reached bands AA and AAA. This demonstrates that automotive datasets must be improved significantly for training future IDS solutions. Based on our observations, we propose the following improvements to dataset authors:

1. Improvement of documentation and better labeling.
2. Inclusion of more recent and more varied datasets and more realistic attacks.
3. Assessment of the overall dataset health before publication, for example using the presented IDS data readiness approach.

Alongside the new data readiness definition we provided some examples for quantifying each deficiency with no or minimal domain knowledge. Some of these could be improved in future work, for example by finding better methods for estimating dataset realism, size and complexity.

ACKNOWLEDGMENTS

This research was supported by the Vinnova FFI project "CyReV: Cyber Resilience for Vehicles" under the grants 2018-05013 and 2019-03071.

REFERENCES

[1] 2021. *UN Regulation No. 155.* Regulation E/ECE/TRANS/505/Rev.3/Add.154. United Nations.

[2] Fahd A Alhaidari and Alia Mohammed Alrehan. 2021. A simulation work for generating a novel dataset to detect distributed denial of service attacks on Vehicular Ad hoc NETwork systems. *International Journal of Distributed Sensor Networks* 17, 3 (2021).

[3] Maytheewat Aramrattana, Jérôme Detournay, Cristofer Englund, Viktor Frimodig, Oscar Uddman Jansson, Tony Larsson, Wojciech Mostowski, Víctor Díez Rodríguez, Thomas Rosenstatter, and Golam Shahanoor. 2018. Team Halmstad Approach to Cooperative Driving in the Grand Cooperative Driving Challenge 2016. *IEEE Transactions on Intelligent Transportation Systems* 19, 4 (2018), 1248–1261.

[4] Stefan Axelsson. 2000. *Intrusion Detection Systems: A Survey and Taxonomy.* Technical Report. Chalmers University of Technology.

[5] Viacheslav Belenko, Vasiliy Krundyshev, and Maxim Kalinin. 2018. Synthetic Datasets Generation for Intrusion Detection in VANET. In *Proceedings of the 11th International Conference on Security of Information and Networks* (Cardiff, United Kingdom) *(SIN '18).* Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages.

[6] Zixiang Bi, Guoai Xu, Guosheng Xu, Miaoqing Tian, Ruobing Jiang, and Sutao Zhang. 2022. Intrusion Detection Method for In-Vehicle CAN Bus Based on Message and Time Transfer Matrix. *Security and Communication Networks* 2022 (07 Mar 2022), 2554280.

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11618–11628.

[8] Laurens A. Castelijns, Yuri Maas, and Joaquin Vanschoren. 2020. The ABC of Data: A Classifying Framework for Data Readiness. In *Machine Learning and Knowledge Discovery in Databases*, Peggy Cellier and Kurt Driessens (Eds.). Springer International Publishing, Cham, 3–16.

[9] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter* 6, 1 (June 2004), 1–6.

[10] Stephen Checkoway, Damon McCoy, Brian Kantor, Danny Anderson, Hovav Shacham, Stefan Savage, Karl Koscher, Alexei Czeskis, Franziska Roesner, and Tadayoshi Kohno. 2011. Comprehensive Experimental Analyses of Automotive Attack Surfaces. In *20th USENIX Security Symposium (USENIX Security 11).* USENIX Association, San Francisco, CA. https://www.usenix.org/conference/usenix-security-11/comprehensive-experimental-analyses-automotive-attack-surfaces

[11] Irina Chiscop, András Gazdag, Joost Bosman, and Gergely Biczók. 2021. Detecting Message Modification Attacks on the CAN Bus with Temporal Convolutional Networks. In *Proceedings of the 7th International Conference on Vehicle Technology and Intelligent Transport Systems.*

[12] Kyong-Tak Cho and Kang G. Shin. 2016. Fingerprinting Electronic Control Units for Vehicle Intrusion Detection. In *25th USENIX Security Symposium (USENIX Security 16).* USENIX Association, Austin, TX, 911–927.

[13] Carlos Garcia Cordero, Emmanouil Vasilomanolakis, Aidmar Wainakh, Max Mühlhäuser, and Simin Nadjm-Tehrani. 2021. On Generating Network Traffic Datasets with Synthetic Attacks for Intrusion Detection. *ACM Transactions on Privacy and Security* 24, 2 (feb 2021), 1–39.

[14] Guillaume Dupont, Jerry Den Hartog, Sandro Etalle, and Alexios Lekidis. 2019. Evaluation Framework for Network Intrusion Detection Systems for In-Vehicle CAN. In *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE).* 1–6.

[15] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2021. Information-Theoretic Measures of Dataset Difficulty. *ArXiv* abs/2110.08420 (2021).

[16] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research* 61 (2018), 863–905.

[17] Ronald Aylmer Fisher. 1988. Iris Flower Data Set. UCI Machine Learning Repository.

[18] Mahsa Foruhandeh, Yanmao Man, Ryan Gerdes, Ming Li, and Thidapat Chantem. 2019. SIMPLE: Single-Frame based Physical Layer Identification for Intrusion Detection and Prevention on In-Vehicle Networks. In *Proceedings of the 35th Annual Computer Security Applications Conference.*

[19] Andre Groll and Christoph Ruland. 2009. Secure and authentic communication on existing in-vehicle networks. In *IEEE Intelligent Vehicles Symposium.* 1093–109.

[20] Mee Lan Han, Byung Il Kwak, and Huy Kang Kim. 2018. Anomaly intrusion detection method for vehicular networks based on survival analysis. *Vehicular Communications* 14 (2018), 52–63.

[21] Markus Hanselmann, Thilo Strauss, Katharina Dormann, and Holger Ulmer. 2019. CANet: An Unsupervised Intrusion Detection System for High Dimensional CAN Bus Data. *CoRR* abs/1906.02492 (2019). arXiv:1906.02492 http://arxiv.org/abs/1906.02492

[22] S. Hettich and S. D. Bay. 1999. KDD Cup 1999 Data. *The UCI KDD Archive* (1999). http://kdd.ics.uci.edu/

[23] Tin Kam Ho and Mitra Basu. 2002. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 3 (2002), 289–300.

[24] Nishat I Mowla and Joakim Rosell. 2021. *V2X and cloud-based Intrusion Detection Mechanisms.* Report CyReV (Cybersecurity for automotive systems in a changing environment) Deliverable 3.1.

[25] Safras Iqbal, Peter Ball, Muhammad H Kamarudin, and Andrew Bradley. 2022. Simulating Malicious Attacks on VANETs for Connected and Autonomous Vehicle Cybersecurity: A Machine Learning Dataset. *arXiv preprint arXiv:2202.07704* (2022).

[26] Mafijul Md. Islam, Aljoscha Lautenbach, Christian Sandberg, and Tomas Olovsson. 2016. A Risk Assessment Framework for Automotive Embedded Systems. In *Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security - CPSS 16.* Association for Computing Machinery (ACM).

[27] K. Strandberg, T. Rosenstatter, R. Jolak, N. Nowdehi, and T. Olovsson. 2021. Resilient Shield: Reinforcing the Resilience of Vehicles Against Security Threats. *IEEE 93rd Vehicle Technology Conference* (2021).

[28] Joseph Kamel, Michael Wolf, Rens W. van der Hei, Arnaud Kaiser, Pascal Urien, and Frank Kargl. 2020. VeReMi Extension: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC).* 1–6.

[29] Adi Karahasanovic, Pierre Kleberger, and Magnus Almgren. 2017. Adapting Threat Modeling Methods for the Automotive Industry. 15th ESCAR Conference, Berlin.

[30] Anthony Kenyon, Lipika Deka, and David Elizondo. 2020. Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets. *Computers & Security* 99 (12 2020), 102022.

[31] Mohammed Khodari, Abhimanyu Rawat, Mikael Asplund, and Andrei Gurtov. 2019. Decentralized Firmware Attestation for In-Vehicle Networks. In *Proceedings of the 5th on Cyber-Physical System Security Workshop (CPSS '19).* ACM Press, 47–56.

[32] B. I. Kwak, J. Woo, and H. K. Kim. 2016. Know your master: Driver profiling-based anti-theft method. In *2016 14th Annual Conference on Privacy, Security and Trust (PST).* 211–218.

[33] Jan Lastinec and Mario Keszeli. 2019. Analysis of Realistic Attack Scenarios in Vehicle Ad-hoc Networks. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS).* 1–6.

[34] Neil D. Lawrence. 2017. Data Readiness Levels. *arXiv:1705.02245* (2017). arXiv:1705.02245 http://arxiv.org/abs/1705.02245

[35] H. Lee, S. H. Jeong, and H. K. Kim. 2017. 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).* 57–5709.

[36] Georg Macher, Harald Sporer, Reinhard Berlach, Eric Armengaud, and Christian Kreiner. 2015. SAHARA: A security-aware hazard and risk analysis method. In *2015 Design, Automation Test in Europe Conference Exhibition (DATE).* 621–624.

[37] Charlie McCarthy, Kevin Harnett, and Art Carter. 2014. *Technical report DOT HS 812 074: Characterization of potential security threats in modern automobiles: a composite modeling approach.* Technical Report. Washington DC.

[38] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 3428–3448.

[39] John McHugh. 2000. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. *ACM Trans. Inf. Syst. Secur.* 3, 4 (nov 2000), 262–294.

[40] Microsoft Corporation. 2005. The STRIDE Threat Model. https://msdn.microsoft.com/en-us/library/ee823878.aspx (Accessed: 2022-06-22).

[41] Charlie Miller and Chris Valasek. 2015. Remote exploitation of an unaltered passenger vehicle. *Black Hat USA* 2015 (2015).

[42] Raymond Muller, Yanmao Man, Z Berkay Celik, Ming Li, and Ryan Gerdes. 2022. DRIVETRUTH: Automated Autonomous Driving Dataset Generation for Security Applications. In *Workshop on Automotive and Autonomous Vehicle Security (AutoSec) 2022.*

[43] Dennis K. Nilsson, Ulf E. Larson, and Erland Jonsson. 2008. Efficient In-Vehicle Delayed Data Authentication Based on Compound Message Authentication Codes. In *Proc. of the 68th IEEE Vehicular Technology Conference* (2008). IEEE, 1–5.

[44] NIST FIPS PUB 199 2004. *NIST FIPS PUB 199 – Standards for Security Categorization of Federal Information and Information Systems.* Standard. National Institute of Standards and Technology.

[45] Rabah Rahal, Abdelaziz Amara Korba, and Nacira Ghoualmi-Zine. 2020. Towards the development of realistic dos dataset for intelligent transportation systems. *Wireless Personal Communications* 115, 2 (2020), 1415–1444.

[46] Robert Bosch GmbH. 2012. *CAN with Flexible Data-Rate - Specification Version 1.0*. Robert Bosch GmbH.

[47] Thomas Rosenstatter, Kim Strandberg, Rodi Jolak, Riccardo Scandariato, and Tomas Olovsson. 2020. REMIND: A Framework for the Resilient Design of Automotive Systems. In *2020 IEEE Secure Development (SecDev)*. 81–95.

[48] Sang Uk Sagong, Xuhang Ying, Radha Poovendran, and Linda Bushnell. 2018. Exploring Attack Surfaces of Voltage-Based Intrusion Detection Systems in Controller Area Networks. In *ESCAR'18*.

[49] E. Seo, H. M. Song, and H. K. Kim. 2018. GIDS: GAN based Intrusion Detection System for In-Vehicle Network. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*. 1–6.

[50] Iman Sharafaldin, Amirhossein Gharib, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Towards a reliable intrusion detection benchmark dataset. *Software Networking* 2018, 1 (2018), 177–200.

[51] Florian Sommer, Jürgen Dürrwang, and Reiner Kriesten. 2019. Survey and Classification of Automotive Security Attacks. *Information* 10, 4 (2019).

[52] Dorsaf Swessi and Hanen Idoudi. 2021. A Comparative Review of Security Threats Datasets for Vehicular Networks. In *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. 746–751.

[53] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. IEEE.

[54] Tencent Keen Security Lab. 2020. Experimental Security Assessment on Lexus Cars. https://keenlab.tencent.com/en/2020/03/30/Tencent-Keen-Security-Lab-Experimental-Security-Assessment-on-Lexus-Cars/. Accessed: 2020-09-15.

[55] U.S. Department of Transportation Federal Highway Administration. 2016. Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Accessed: 2022-09-27: from http://doi.org/10.21949/1504477.

[56] U.S. Department of Transportation Federal Highway Administration (FHWA). 2022. Next Generation Simulation (NGSIM). https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm Accessed: 2022-03-11.

[57] Rens W. van der Heijden, Thomas Lukaseder, and Frank Kargl. 2018. VeReMi: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs. In *Security and Privacy in Communication Networks*, Raheem Beyah, Bing Chang, Yingjiu Li, and Sencun Zhu (Eds.). Springer International Publishing, Cham, 318–337.

[58] Miki E. Verma, Michael D. Iannacone, Robert A. Bridges, Samuel C. Hollifield, Bill Kay, and Frank L. Combs. 2020. ROAD: The Real ORNL Automotive Dynamometer Controller Area Network Intrusion Detection Dataset (with a comprehensive CAN IDS dataset survey & guide). arXiv:2012.14600 [cs.CR]

[59] W. Wu, R. Li, G. Xie, J. An, Y. Bai, J. Zhou, and K. Li. 2020. A Survey of Intrusion Detection for In-Vehicle Networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 3 (2020), 919–933.

[60] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A Theory of Usable Information Under Computational Constraints. *ArXiv* abs/2002.10689 (2020).

[61] M. Yan, J. Li, and G. Harpak. 2020. Security Research on Mercedes-Benz: From Hardware to Car Control. https://i.blackhat.com/USA-20/Thursday/us-20-Yan-Security-Research-On-Mercedes-Benz-From-Hardware-To-Car-Control.pdf. Accessed: 2020-09-15.