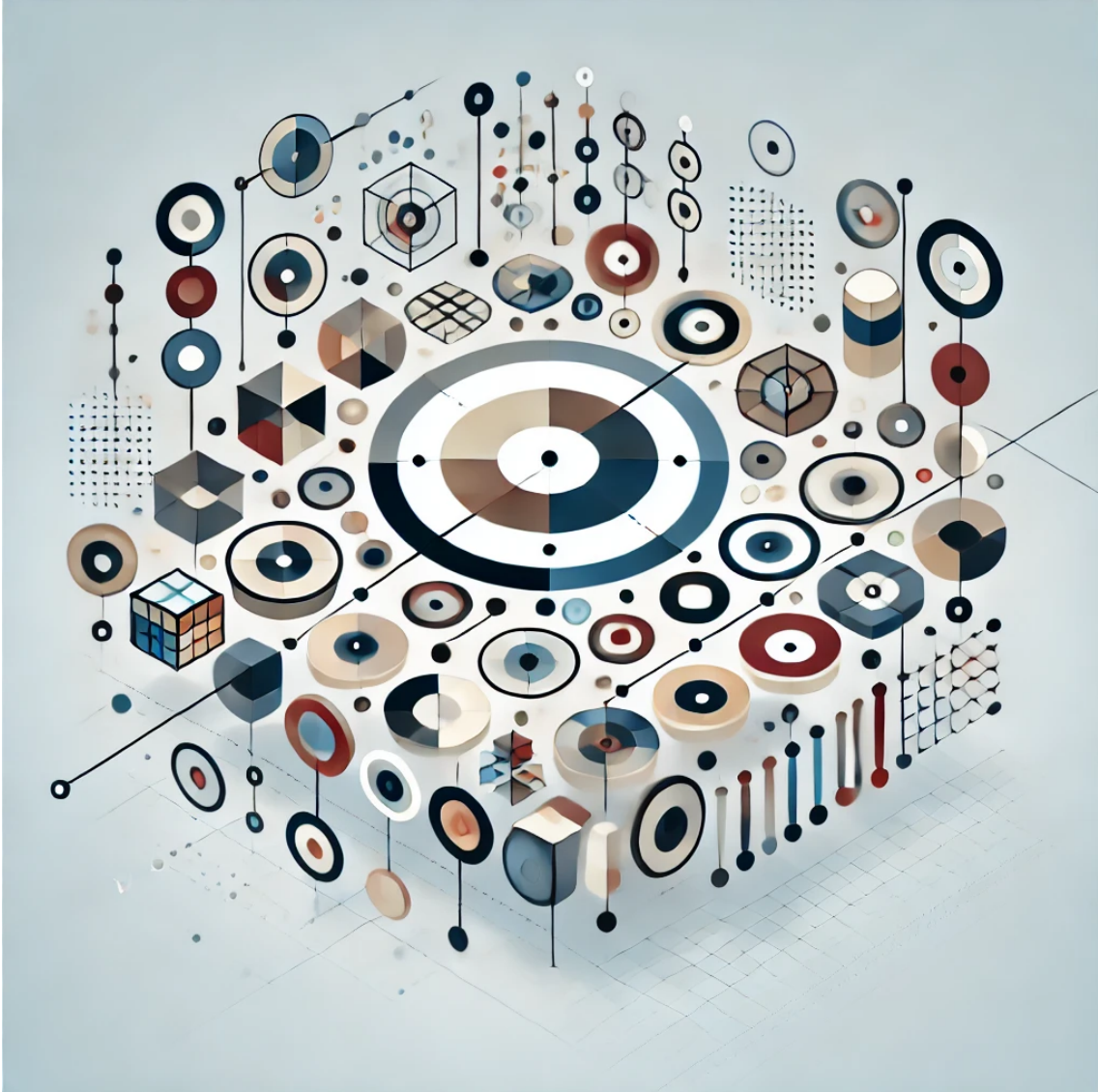


RISE

DIGITAL SYSTEMS
APPLIED DIGITALIZATION



Data Quality Evaluation of CAN and Automotive Ethernet Datasets

Nishat I Mowla

RISE Report : 2025:43

Data Quality Evaluation of CAN and Automotive Ethernet Datasets

Nishat I Mowla

Reviewers: Niclas Ericsson (RISE), Alireza Dehlaghi Ghadim (RISE),
Sarder Fakhru'l Abedin (Mid Sweden University)

Abstract

Data Quality Evaluation of CAN and Automotive Ethernet Datasets

This report evaluates the data quality of two standard publicly available automotive intrusion detection system datasets: the SOME/IP Attack Dataset and the Survival Analysis Dataset (SAD). The data quality evaluation is guided by the data quality model of the comprehensive ISO/IEC 5259 data quality standard series and incorporates domain specific requirements and data usage context relevant to automotive networks. The automotive network data quality requirements are aligned with the AI Act Article 10 provisions on data and data governance requirements, which emphasizes the importance of trustworthy data subject for AI model training. The report presents key findings and reflections on both datasets to enhance understanding, ensure compliance, and support their adoption in the development and validation of AI/ML-based automotive intrusion detection systems.

Key words: Automotive datasets, Automotive intrusion detection data, AI Act, ISO/IEC 5259 series, data quality

Cover illustration: Created with OpenAI's GPT-4o¹. The abstract was included as the prompt.

Acknowledgement: This work is supported by the Vinnova INTelligent sEcuRity SoluTIons for Connected vEhicles (INTERSTICE) project (reference number: 2024-00661). This work is also partially supported by the EU project Citcom.AI, one of the EU's four AI TEFs (Testing and Experimental Facilities for Smart and Sustainable Cities and Societies).

RISE Research Institutes of Sweden AB
RISE Report : 2025:43
ISBN: 978-91-90036-30-3

¹ <https://openai.com/index/gpt-4o>

Content

Abstract	2
Content	3
1 Standardized data quality evaluation	4
1.1 High Level AI Act data quality requirements	4
1.2 Data quality standard ISO/IEC 5259 series.....	5
2 Data quality evaluation in Automotive IDS	6
2.1 Automotive IDS datasets	6
2.2 Automotive IDS data quality needs.....	7
2.3 Data quality model on automotive IDS datasets	8
3 Performance Evaluation	9
3.1 Accuracy	9
3.2 Completeness.....	11
3.3 Consistency	14
3.4 Diversity	18
3.5 Credibility	20
3.6 Currentness.....	21
4 Observations and conclusion	22

1 Standardized data quality evaluation

1.1 High Level AI Act data quality requirements

The European Union's AI Act, specifically Article 10², mandates that data used for training, testing, and validating AI systems must adhere to high-quality standards to ensure fairness, reliability, and accountability. This includes ensuring that datasets are **free of errors, complete, relevant, and sufficiently representative**, for the intended purpose of the AI system: 1) *Free of Errors*: Errors in data, such as incorrect labels, duplicate records, or anomalies, can distort the performance of AI models. Rigorous validation and cleansing processes should be applied to detect and rectify such issues, 2) *Completeness*: Datasets must include all necessary and meaningful data points to accurately reflect the problem domain. Missing or incomplete data can lead to biased AI models or incorrect conclusions, 3) *Relevance*: Data should be directly related to the AI system's intended purpose. Irrelevant or extraneous data may introduce noise, inefficiencies, or inaccuracies, ultimately affecting model performance and fairness, 4) *Representativeness*: To mitigate bias and promote fairness, datasets must represent the diversity and characteristics of the population or domain being modeled. This includes ensuring balanced representation across demographic, geographic, and contextual variables where applicable. Addressing these gaps is essential for compliance.

According to European Commission's JRC (Joint Research Center) Technical Report³, the ISO/IEC 5259 series primarily addresses data-related aspects concerning analytics and machine learning, with part 3 specifically covering AI Act requirements related to data governance. When it comes to data quality, part 2 of the series offers an extensive list of quality attributes, including those most pertinent to the AI Act. However, further implementation measures are necessary to facilitate the *appropriate selection and prioritization* of these attributes in accordance with the risks outlined by the AI Act. Currently, the standards within this series adopt a broad perspective, defining data quality in terms of its ability to meet an organization's specific requirements.

In this report, we follow the guidelines of data quality from the ISO/IEC 5259 series while automotive network domain knowledge is incorporated to select, prioritize, and perform further evaluation of the quality of the datasets considered in this report.

² Regulation (EU) 2024/1689 of the European Parliament and of the Council of the 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), <https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>.

³ Analysis of the preliminary AI standardisation work plan in support of the AI Act. Available at <https://publications.jrc.ec.europa.eu/repository/handle/JRC132833>

1.2 Data quality standard ISO/IEC 5259 series

As discussed in the previous section, the data quality requirements set out in the AI Act are effectively supported by the ISO/IEC 5259 series on data quality for analytics and machine learning. This series provides a comprehensive framework for evaluating and managing data quality within AI systems. Part 2 of the series offers a detailed catalogue of quality attributes that align closely with the AI Act's requirements. These attributes serve as a foundation for assessing data quality in terms of *relevance*, *completeness*, *representativeness*, and *freedom from errors*. Although the ISO/IEC 5259 series define data quality broadly as "data meeting the organization's requirements," additional implementation guidelines are necessary. These guidelines should prioritize and select quality attributes based on the domain knowledge associated with the AI system's intended purpose. To support this, ISO/IEC 5259-2 introduces a structured data quality model that facilitates the specification and evaluation of data quality requirements.

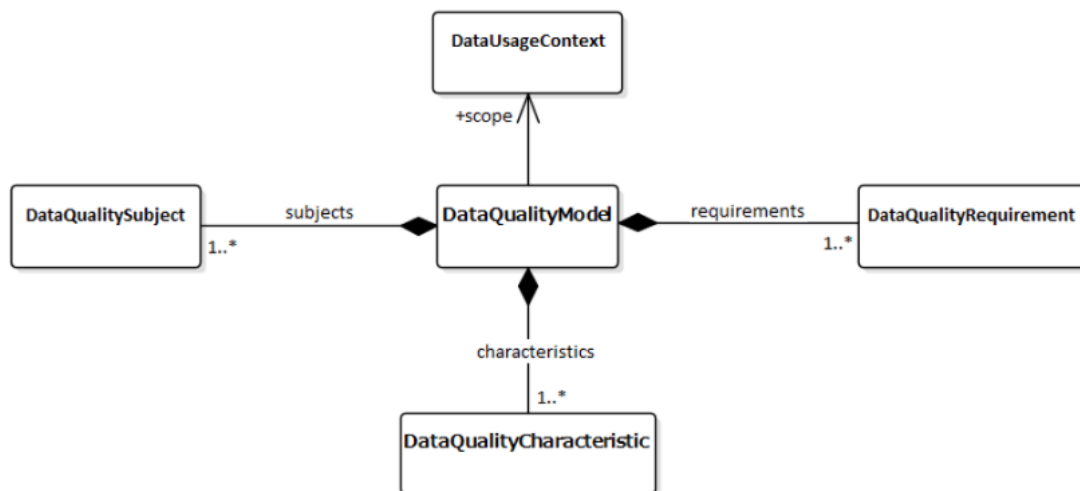


Figure 1. ISO/IEC 5259-2 data quality model.⁴

The ISO/IEC 5259-2 *data quality model*⁵ is composed of key elements: *data quality subjects*, which refer to entities affected by data quality; *data quality characteristics*, which encompass attributes such as accuracy, completeness, and precision; and *data quality requirements*, which define the specific properties of data along with acceptance criteria tailored to its intended use. These elements are designed to align with the application *context*, particularly in analytics and machine learning tasks, such as training neural networks to predict product sales based on marketing strategies. Represented through a UML (Unified Modeling Language) diagram as will be further discussed and elaborated in the next section, the model emphasizes the critical role of context in defining and achieving data quality goals. It provides organizations with the ability to select appropriate data quality attributes and measures to meet specific needs. Furthermore, data

⁴ ISO/IEC 5259-1:2024 Artificial intelligence — Data quality for analytics and machine learning (ML). Available at <https://www.iso.org/standard/81088.html>

⁵ N. I. Mowla, 'A Guide to Data Quality Testing for AI Applications based on Standards', RISE Research Institutes of Sweden, 2024.

quality requirements within the framework can be expressed in quantitative, qualitative, or context-dependent terms, ensuring that the data is suitable for specific applications.

2 Data quality evaluation in Automotive IDS

2.1 Automotive IDS datasets

Automotive Intrusion Detection System (IDS) datasets are essential for developing and evaluating security mechanisms to protect in-vehicle networks against cyber threats. These datasets typically include network traffic data collected from automotive communication protocols such as Controller Area Network (CAN) and more recently Automotive Ethernet, e.g. SOME/IP (Scalable service-Oriented MiddlewarE over IP), and others. In general, these datasets contain normal and attack scenarios, helping researchers analyze intrusion patterns, detect anomalies, and improve cybersecurity measures for modern vehicles. For this data quality report, we investigate two standard publicly available automotive IDS datasets, i.e., SOME/IP dataset as an Automotive Ethernet data source and the Survival Analysis Dataset as the CAN data source.

The SOME/IP IDS dataset⁶ contains network traffic data related to the SOME/IP protocol, which is widely used in automotive Ethernet communication. This dataset includes labeled data samples capturing both normal and anomalous traffic, making it useful for training and evaluating intrusion detection models.

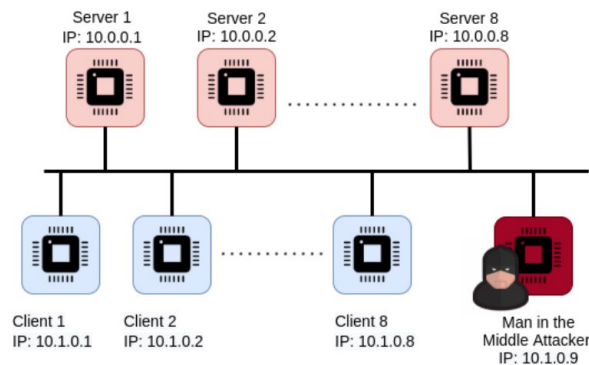


Figure 2. Different SOME/IP clients and servers exchanging SOME/IP services over Automotive Ethernet bus where a Client ECU is being compromised by an MITM Attacker⁶.

Attack scenarios related to the vulnerabilities in the SOME/IP protocol used in automotive Ethernet networks include several threats. One specific attack is *Requests without Response*, where requests are sent without receiving any responses, potentially

⁶ Alkhatib, N., Ghauch, H., & Danger, J. L. (2021, October). SOME/IP intrusion detection using deep learning-based sequential models in automotive ethernet networks. In 2021 IEEE 12th annual information technology, electronics and mobile communication conference (IEMCON) (pp. 0954-0962). IEEE.

indicating message interception or loss. Another threat, *Response without Request*, involves sending responses when no corresponding requests have been made, suggesting unsolicited or rogue responses. The *Error on Error* scenario occurs when error messages are sent in response to other error messages, which violates protocol standards. Similarly, *Error on Event* involves sending error messages in reaction to event notifications, an action that should not trigger errors under normal protocol operations.

The Survival Analysis Dataset⁷ provides network intrusion data that supports survival analysis, a statistical technique used to estimate the time until a cyber-attack occurs. This dataset includes timestamps and event-based information, helping researchers analyze the persistence and timing of attacks in network environments.

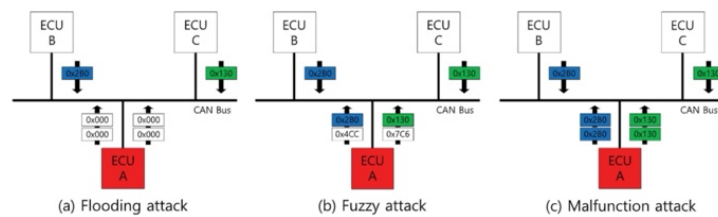


Figure 3. Flooding, fuzzy and malfunction attacks in CAN⁷.

The Survival Analysis Dataset (SAD) consists of CAN-bus data collected from three different vehicles: Hyundai Sonata, Kia Soul, and Chevrolet Spark. The dataset includes various attack scenarios such as *Flooding*, *Fuzzy*, and *Malfunction*, which can significantly disrupt in-vehicle functions. The Flooding attack aims to overwhelm the CAN bus by continuously occupying its resources, maintaining a dominant status, and preventing legitimate ECU messages from being transmitted. This type of attack can effectively disable critical vehicle functions by monopolizing the communication channel. The Fuzzy attack involves the iterative injection of random CAN packets into the network. By introducing unpredictable and malformed messages, this attack can lead to erratic system behavior, cause unexpected vehicle responses, or crash ECUs that are unable to handle anomalous data. The Malfunction attack specifically targets a selected CAN ID extracted from the vehicle. This attack manipulates the data field while simultaneously injecting randomly selected CAN IDs. By modifying the 8-byte data field with values set to 00 or other arbitrary numbers, the attack forces the vehicle into abnormal behavior, leading to unintended operations, performance degradation, or even critical system failures. These attack types have the potential to not only impair the normal operation of the vehicle systems but also escalate the severity of an attack and increase the extent of the resulting impact.

2.2 Automotive IDS data quality needs

Automotive IDS data quality is crucial for ensuring the accuracy and reliability of intrusion detection systems in vehicles. High-quality data must exhibit characteristics

⁷ Han, M. L., Kwak, B. I., & Kim, H. K. (2018). Anomaly intrusion detection method for vehicular networks based on survival analysis. *Vehicular communications*, 14, 52-63.

such as completeness of attack coverage and essential features, consistency of formats, semantic and syntactic accuracy fit to the purpose, and relevance to effectively detect and mitigate cybersecurity threats. Data governance practices, including proper labeling and validation, are essential to minimize noise and inconsistencies that could lead to false positives or negatives. Furthermore, the selection and prioritization of quality attributes should align with the specific needs and challenges posed by automotive environments, such as real-time processing requirements and the dynamic nature of vehicular networks. Ensuring robust data quality is fundamental for enhancing the performance and trustworthiness of automotive IDS solutions.

2.3 Data quality model on automotive IDS datasets

As mentioned earlier, ISO/IEC 5259-2 discusses the data quality model that serves as the foundation for understanding and managing data quality attributes. This standardized data quality model links a data quality context to specific requirements, characteristics, and entities affected by the quality of the data. In this report, we have mapped the automotive intrusion detection system (IDS) dataset's requirements, selected and prioritized characteristics, data quality subject and data usage context on this standardized data quality model, to perform the data quality evaluation.

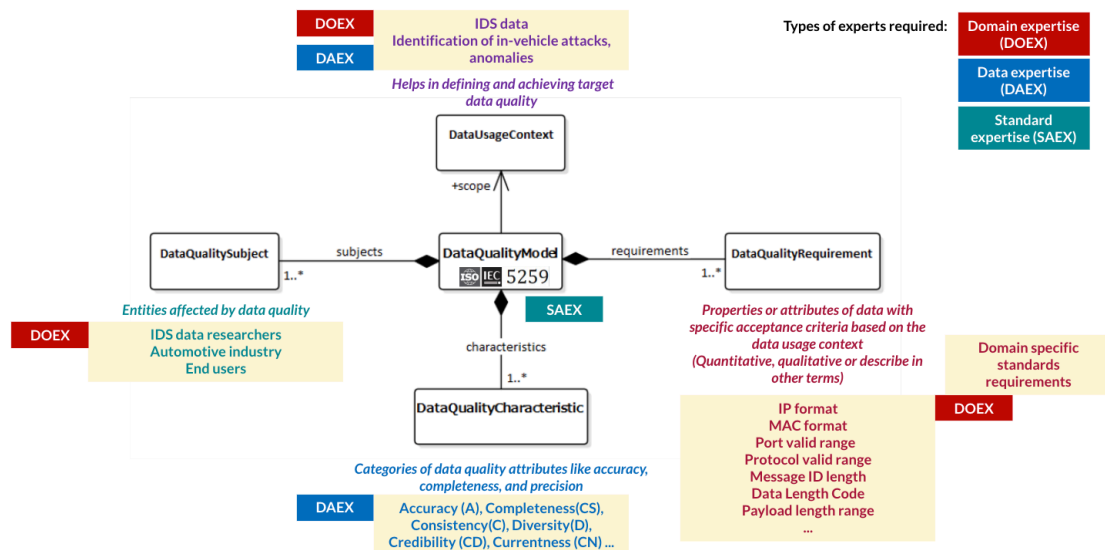


Figure 4. Data quality model for automotive in-vehicle IDS data.

Figure 4 illustrates the application of the ISO/IEC 5259 series data quality model to assess and improve the data quality of in-vehicle Intrusion Detection System (IDS) data within the automotive domain. The model helps to assess and combine: 1) *data quality requirements*: defines the properties or attributes of data with specific acceptance criteria tailored to the data usage context. These requirements may be quantitative, qualitative, or descriptive and include IP (Internet Protocol) format, MAC (Media Access Control) format, port valid range, protocol valid range, message ID length, payload length range. These requirements are guided by domain-specific standards to ensure IDS data meets the necessary benchmarks for reliability and functionality. 2) *data quality characteristics*: categories of data quality attributes that are critical for IDS data, such as accuracy (A),

completeness (CS), consistency (C), diversity (D), credibility (CD), currentness (CN). These attributes ensure the IDS data is robust and reliable for its intended applications, 3) *data quality subject*: represents the entities affected by the data quality. For example, IDS data researchers focused on analyzing and understanding the data for IDS effectiveness, automotive industry concerned with integrating IDS data for vehicle safety and security, and end users ultimately affected by the quality of IDS outputs, such as alerts or recommendations, and 4) *data usage context*: defines the scope of the data usage, such as identifying in-vehicle attacks or detecting anomalies in automotive systems. This helps establish clear expectations and objectives for data quality requirements.

We further refine the data quality model with the different expertise required to ensure effective data quality evaluation for IDS systems. Figure 4 highlights three types of essential expertise for implementing the data quality model effectively: 1) *Domain Expertise (DOEX)*: Knowledge about IDS data, in-vehicle attacks, and anomalies, 2) *Data Expertise (DAEX)*: Understanding of data quality attributes and methodologies for achieving the targeted data quality, 3) *Standard Expertise (SAEX)*: Familiarity with ISO/IEC 5259 and other domain-specific standards to ensure compliance.

The model bridges the gap between theoretical data quality management and its practical implementation in the automotive domain. By aligning IDS data with specific quality requirements and leveraging domain, data, and standards expertise, this approach ensures that the IDS should be able to effectively identify in-vehicle attacks and anomalies, thereby enhancing vehicle cybersecurity and user trust.

For the data quality evaluation of automotive intrusion detection systems, we focus on assessing and prioritizing the inherent data quality characteristics—accuracy, completeness, consistency, credibility, and currentness—along with the additional characteristic of diversity. The rationale behind selecting these characteristics is based on an analysis of the use case specific requirements, which highlights that all inherent data quality characteristics are equally essential for the data quality evaluation. Besides, the evaluation of diversity (that also includes distribution of data records) are additionally required as they play a crucial role in enhancing the performance of automotive intrusion detection systems. In particular, diversity ensures that the dataset captures a wide range of possible scenarios, reducing bias and improving the system's robustness against various attack patterns. We do not perform the evaluation of the system-dependent data quality characteristics as it is assumed that these requirements have been addressed during data collection and processing stages. Therefore, their assessment falls outside the scope of this evaluation.

3 Performance Evaluation

3.1 Accuracy

Accuracy, defined as the closeness of agreement between an observed value and the true or reference value, is a critical data quality characteristic for automotive intrusion detection systems (IDS). The evaluation of accuracy can be categorized into different dimensions as defined by ISO/IEC 5259 series:

- 1) *Syntactic accuracy* refers to the closeness of data values to a predefined set within a specific domain, ensuring consistency in formatting and structure,

- 2) *Semantic accuracy*, on the other hand, assesses how well the data values reflect their intended meaning in a given context,
- 3) *Data accuracy assurance* involves measuring the coverage of accurate data within a dataset, ensuring comprehensive validation,
- 4) *Risk of dataset inaccuracy* is determined by the number of outliers present, which could indicate potential data reliability issues and,
- 5) *Data accuracy range* evaluates whether the data values fall within a required interval, ensuring consistency within acceptable boundaries.

Moreover, ensuring accuracy involves addressing both domain and data-related concerns. One such key aspect is *automotive dataset realism*, which emphasizes the need for data to be recorded in realistic settings using diverse sources where applicable. This enhances the dataset's applicability to real-world scenarios to improve the effectiveness of IDS models. Another important factor is *transformation and anonymization*, where modifications to the data, such as anonymization techniques, may impact its real-world applicability and introduce challenges in achieving IDS accuracy.

Table 1. SOME/IP Accuracy evaluation.

SOME/IP Accuracy	Score	Explanation
Syntactic Accuracy	100%	All data items syntactically accurate.
Semantic Accuracy	100%	All data items semantically accurate.
Data Accuracy Assurance	100%	All data items measured for accuracy.
Risk of Dataset Inaccuracy	0%	No data item with outlier values.
Data Accuracy Range	100%	E.g., No values outside required interval for errorRate . Expected range: 0 to 1

Table 2. SAD Accuracy evaluation.

SAD Accuracy	Score	Explanation
Syntactic Accuracy	100%	All data items syntactically accurate.
Semantic Accuracy	100%	All data items semantically accurate.
Data Accuracy Assurance	100%	All data items measured for accuracy.
Risk of Dataset Inaccuracy	0%	No data item with outlier values.
Data Accuracy Range	100%	E.g., No values outside expected interval for DLC. Expected range 0 to 8

The evaluation of the SOME/IP and SAD datasets was conducted across multiple accuracy metrics, including syntactic accuracy, semantic accuracy, data accuracy assurance, risk of dataset inaccuracy, and data accuracy range as shown in Table 1 and Table 2. The results indicate that the SAD dataset demonstrates 100% accuracy in all accuracy metrics. Data accuracy range is shown for DLC (Data Length Code) which scored 100%, with no values found outside the expected interval for DLC (i.e., 0 to 8 byte). In the case of the SOME/IP dataset, the syntactic accuracy, semantic accuracy, data accuracy assurance was also 100% and risk of dataset inaccuracy was 0% as no significant outliers were found and values were all within required intervals as shown through data accuracy range. For example, the data accuracy range for errorRate was set to 0 to 1 as it is a probability that the request is not answered with a response. All values were found

within the 0 to 1 range. These findings ensure that in general both datasets exhibit essential data quality standards.

The SOME/IP dataset is generated by using a specialized tool⁸ that simulates both normal and attack scenarios based on AUTOSAR specifications, ensuring realism and applicability to real-world automotive environments. The dataset is generated based on specific configuration parameters that define its structure and characteristics. These parameters include the number and types of devices involved, consisting of eight servers, eight clients, and one attacker. Additionally, the dataset specifies three services that are offered and requested. For each combination of a client, a service, and a method within that service, 50 packets are generated, ensuring a structured distribution of network traffic. Attack scenarios, such as "Error On Error," "Error On Event," "Missing Request," and "Missing Response," are implemented to evaluate intrusion detection system (IDS) performance under realistic conditions. Additionally, response times for attackers are configured with minimum and maximum values of 1 ms and 3 ms, respectively, to simulate the variations normally seen in real systems. The resulting dataset is stored in an output file named output.pcap for further analysis. To address the challenges associated with variable-length sequence prediction in IDS, the dataset undergoes a transformation process to ensure uniform sequence lengths. Each sequence contains a maximum of 60 packets, with sequences padded with zeros if they contain fewer packets. This transformation facilitates consistent analysis and model training by maintaining a fixed input length. No anonymization techniques are expected in the dataset since the dataset is generated by a simulator.

The SAD dataset demonstrates high level of realism by generating packets using real vehicles, providing authentic data that closely reflects real-world automotive environments. This approach enhances the applicability of the dataset for evaluating intrusion detection systems (IDS) representing realistic driving and operational conditions. The use of actual vehicle data helps capture the inherent complexities and variabilities present in automotive networks, making the dataset suitable for practical IDS development and testing. Regarding transformation and anonymization, no specific techniques have been mentioned in the dataset documentation. This implies that the dataset retains its original structure and content without modifications or masking of sensitive information. The absence of transformations ensures that the data reflects its raw state, which can be beneficial for accurately assessing IDS performance in detecting genuine network anomalies and attack scenarios.

3.2 Completeness

Completeness, defined as the presence or absence of features or attributes in the dataset, is a crucial data quality characteristic for automotive intrusion detection systems (IDS). It encompasses both data omission (missing or incomplete data) and data commission (extra or irrelevant data), ensuring that the dataset sufficiently supports intended analyses. The evaluation of completeness can be categorized into different dimensions as defined by ISO/IEC 5259:

⁸ SOME/IP Generator, "SOME/IP Generator Documentation," Available: <https://some-ip-generator.readthedocs.io/en/latest/>, Accessed: January 20, 2025.

- 1) *Value completeness* refers to the ratio of data items with non-null values, ensuring that missing data is minimized.
- 2) *Value occurrence completeness* evaluates whether the number of occurrences of a given data value aligns with expected distributions, ensuring that certain events or features are not underrepresented.
- 3) *Feature completeness* measures whether all expected features contain meaningful data, ensuring that crucial attributes are not omitted from the dataset.
- 4) *Record completeness* assesses the presence of empty records, ensuring that all dataset entries contain valid information.
- 5) *Label completeness* refers to the presence of correct and fully assigned labels for each data instance, ensuring that no entries remain unlabeled or ambiguously classified.

Ensuring completeness requires addressing both domain-specific and data-related concerns. One critical factor is dataset size, which determines whether the dataset includes a sufficiently large number of samples and relevant features to support comprehensive IDS evaluation. Another key aspect is *attack completeness*, ensuring that a sufficient number of attack instances are included, covering diverse scenarios to enable effective model training and validation. Additionally, OSI (Open Systems Interconnection) layer representation plays a fundamental role in IDS dataset completeness, as it ensures that data is captured from appropriate network layers relevant to intrusion detection.

Table 3: SOME/IP completeness evaluation.

SOME/IP Completeness	Score	Explanation
Value Completeness	100%	All data items with no null values
Value Occurrence Completeness	55.27%	Label 0 has 5659 training values Label 1 has 315 training values Total observed occurrences: 5974 Number of unique values: 2 Expected occurrences per value: 2987.00 Sum of min (observed, expected): 3302.0 Final Value Occurrence Completeness: 55.27% <100%: Imbalanced distribution (some values appear more or less frequently than others).
Feature Completeness	100%	All data item having a feature value not null
Record Completeness	100%	All data records with no empty data item
Label Completeness	100%	All samples labelled and no incompletely labelled data item
Number of Data Points	Train: 5,974, Test: 6,091	
Number of Features	35	
Number of Classes	2	Each file has two classes

Names of Classes	[0, 1]	Each file has normal data labelled to 0 and attack data which is labelled to 1
OSI Layer Coverage	5-7	
Attack Completeness	{'Spoofing': 1, 'Tampering': 1, 'Repudiation': 0, 'Information Disclosure': 1, 'DoS': 1, 'Elevation of Privilege': 0}	

Table 4. SAD completeness evaluation.

SAD Completeness	Score	Explanation
Value Completeness	100%	All data items with no null values
Value Occurrence Completeness	60.56%	Label 0 has 1552526 values and Label 1 has 183314 values Total observed occurrences: 1735840 Number of unique values: 2 Expected occurrences per value: 867920.00 Sum of min (observed, expected): 1051234.0 Final Value Occurrence Completeness: 60.56% <100%: Imbalanced distribution (some values appear more or less frequently than others).
Feature Completeness	100%	All data item having a feature value not null
Record Completeness	100%	All data records with no empty data item
Label Completeness	100%	All samples labelled and no incompletely labelled data item
Number of Data Points	1,735,840	
Number of Features	14	
Number of Classes	2	Each file has two classes
Names of Classes	[0, 1]	Each file has normal data labelled to 0 and attack data which is labelled to 1
OSI Layer Coverage	1-2	
Attack Completeness	{'Spoofing': 1, 'Tampering': 1, 'Repudiation': 0, 'Information Disclosure': 0, 'DoS': 1, 'Elevation of Privilege': 0}	

The SOME/IP dataset ensures completeness by simulating attack scenarios targeting critical vulnerabilities in the SOME/IP protocol, which is crucial for evaluating intrusion detection in automotive Ethernet communication. Its OSI layer representation focuses on the OSI layers 5 to 7, which are the primary layers affected by network-based SOME/IP attacks. By covering these layers, the dataset ensures an effective and structured representation of intrusion attempts, making it relevant for automotive IDS research. The dataset contains 5,974 training and 6,091 testing data points each with 35 features with

no null values. The requests without response attack occurs when messages are intercepted, blocked, or lost, disrupting communication or exposing data leading to the possibility of denial of service and information disclosure. The response without request attack involves sending rogue responses to mislead the system with false data which maps to spoofing and tampering-based attack. The error on error attack injects error messages in response to other errors, violating protocol standards and potentially destabilizing the network leading to tampering and denial of service (DoS) attacks. Similarly, the error on event attack, also classified as tampering and DoS, sends false error messages in response to legitimate event notifications, overwhelming the system with unnecessary disruptions. While the dataset effectively represents network-layer protocol attacks, it does not explicitly cover elevation of privilege attacks, as these typically exploit authentication mechanisms rather than message-level vulnerabilities. This dataset is well-suited for evaluating IDS models in the context of network-based threats in automotive Ethernet systems.

The SAD dataset demonstrates strong completeness by including data from multiple vehicles (Hyundai Sonata, Kia Soul, and Chevrolet Spark), ensuring coverage in vehicular sources. The dataset captures in-vehicle functions and lower-layer CAN-bus messages, strengthening its OSI layer representation (i.e., layer 1 to 2) and making it highly suitable for comprehensive IDS analysis. The dataset contains 1735840 data points each with 14 features with no null values. In particular, the dataset primarily covers DoS, spoofing, and tampering attacks, making it highly relevant for evaluating IDS models against injection-based threats in CAN networks. The flooding attack, classified as a DoS attack, disrupts normal ECU communication by injecting excessive messages with CAN ID 0x000 (hexadecimal representation of an 11-bit identifier), overwhelming the CAN bus and blocking legitimate messages. The fuzzy attack encompasses spoofing, tampering, and DoS, where randomly generated CAN packets are injected at high frequency, altering both CAN IDs and data fields. This can trick the vehicle into responding to false messages (spoofing), modify system behavior (tampering), or flood the bus with invalid data, leading to DoS. The malfunction attack is categorized under spoofing and tampering, as it involves injecting manipulated data into selected CAN messages from real vehicles. The malfunction attack targets a selected CAN ID from among the extractable CAN IDs of a certain vehicle, causing unintended vehicle responses. Additionally, the dataset includes attack-free states, representing normal driving conditions as a baseline for comparison. The dataset lacks explicit privacy-related or privilege escalation scenarios.

Overall, both datasets exhibit a high level of completeness in terms of dataset size, attack coverage, and OSI layer representation, ensuring their suitability for intrusion detection research in the automotive domain. The value occurrence completeness in both the datasets is lower showing an imbalanced distribution as explained in Table 3 and Table 4. We look into the distributions and diversity in the section 3.4.

3.3 Consistency

Consistency refers to the degree to which data adheres to logical rules and relationships, ensuring that datasets remain structured, error-free, and reliable for analysis. It is evaluated across format correctness, file format consistency, data record consistency, distribution of data values, data format consistency, and semantic consistency. Format correctness ensures that data is stored without errors and does not contain duplicate

records, maintaining the integrity of the dataset. File format consistency guarantees that each data type follows a standardized structure, making the dataset easily interpretable across different systems. The datasets are presented in appropriate formats, such as PCAP (Packet Capture), CSV (Comma-Separated Values), TXT (Text file), or JSON (JavaScript Object Notation), to support both human-readable and machine-readable processing.

- 1) *Data record consistency* is assessed by identifying and minimizing duplicate records, ensuring that redundant or conflicting entries do not compromise the dataset's reliability.
- 2) *Distribution of data values* examines the statistical distribution of feature values, ensuring that attributes maintain expected variations and do not introduce biases or inconsistencies that could affect machine learning models.
- 3) *Data format consistency* focuses on ensuring that properties remain consistent across different data files for format validation.
- 4) *Semantic consistency* verifies that data values conform to predefined semantic rules, ensuring that attributes reflect their intended meaning correctly.

Both the datasets have 0 duplicate records and 100% symantic accuracy. SOME/IP has 94.12% data format consistency as some values such as `client_min` and `client_max` are identified as floating values where the rule was set to be integers. For the Survival Analysis Dataset (SAD), the data format consistency is 100% and file format correctness is maintained as the dataset is stored in txt format without notable issues, ensuring it remains machine-readable. File format consistency is ensured, as all files adhere to the format standard, making data processing and integration straightforward as shown in Table 6.

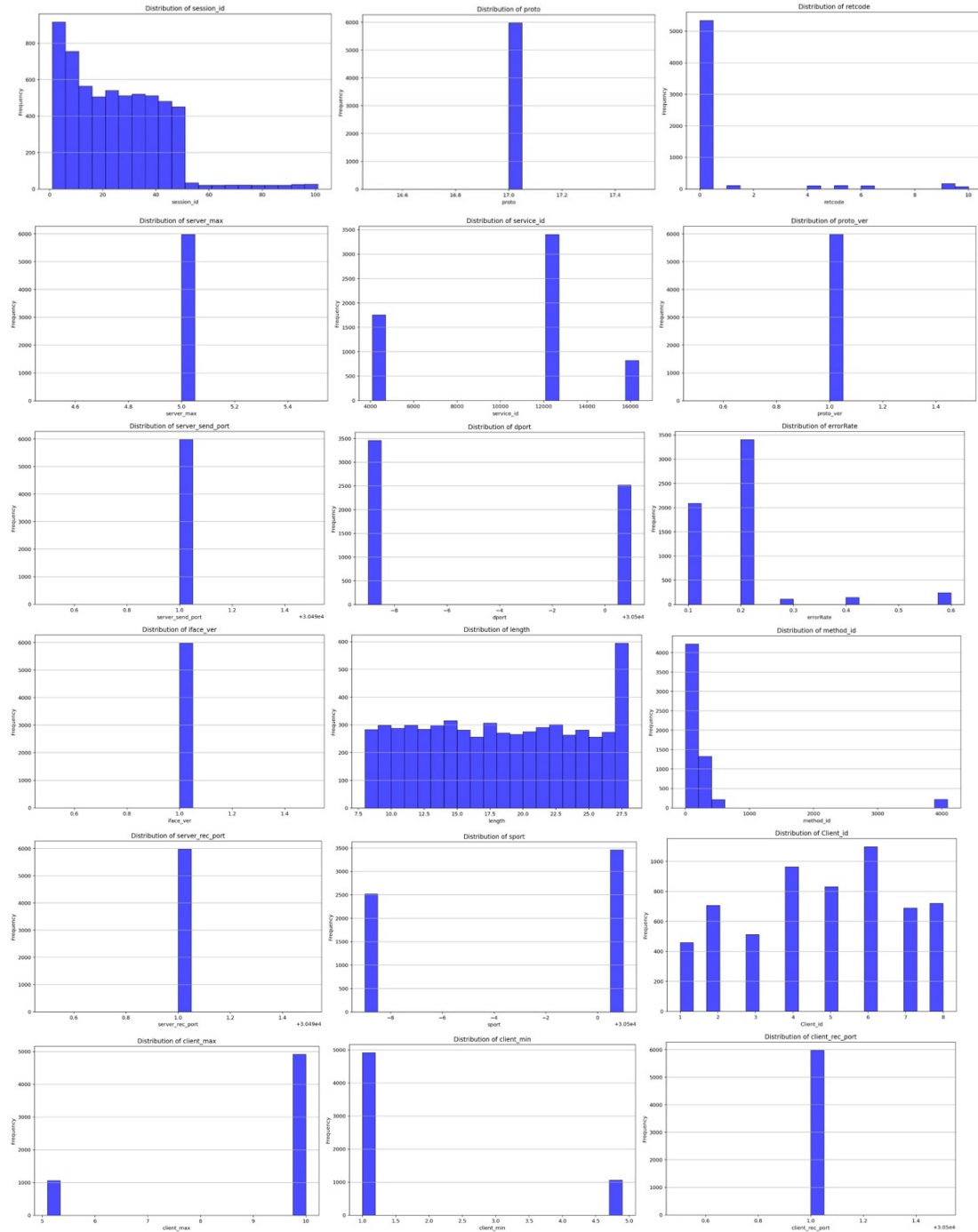
Table 5. SOME/IP consistency.

SOME/IP Consistency	Score	Explanation
Data record consistency	0%	Number of duplicate records out of total records is 0% in the dataset
Data format consistency	100%	All data items with consistent data format
File format consistency		4 csv files
Symantic consistency	100	All data items are symantically correct

Table 6. SAD consistency

SAD Consistency	Score	Explanation
Data record consistency	0%	Number of duplicate records out of total records is 0% in the dataset
Data format consistency	100%	All data items with consistent data format
File format consistency		12 txt files
Symantic consistency	100%	All data items are symantically correct

For the SOME/IP dataset, format correctness is upheld through data simulation, which ensures adherence to AUTOSAR specifications⁹. This guarantees that the dataset follows the expected structure for automotive Ethernet networks. File format consistency is preserved as the dataset is provided in CSV formats as shown in Table 5, ensuring compatibility with standard tools used in automotive Ethernet-based intrusion detection systems (IDS). Data record consistency is maintained by minimizing duplicate records.



⁹ AUTOSAR. AUTOSAR Classic Platform 24-11: Specification of Ethernet. AUTOSAR Release R24-11, 2023. [Online]. Available: https://www.autosar.org/fileadmin/standards/R24-11/CP/AUTOSAR_CP_RS_Ethernet.pdf

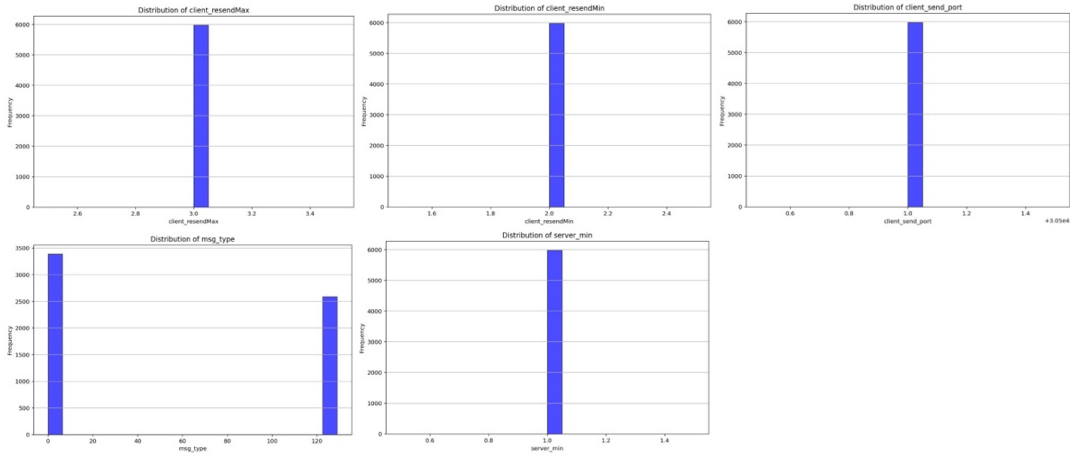


Fig. 2 SOME/IP data feature distribution.

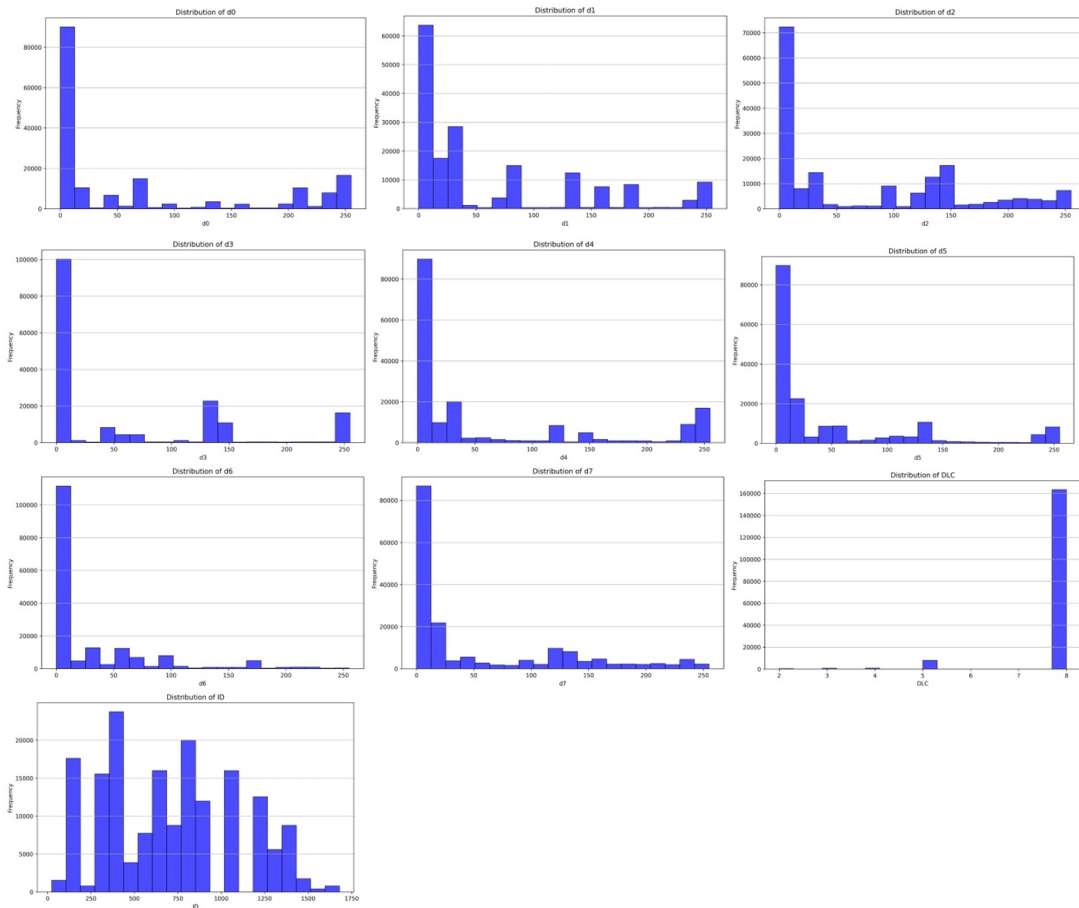


Fig. 3 SAD data feature distribution.

Fig. 2 and Fig. 3 show the feature distribution for the SOME/IP and SAD datasets. The distribution of data values ensures attributes exhibit expected variations without biases that could impact model training. Data format consistency, following ISO/IEC 25024¹⁰ standard, ensures uniform data representation across files. Lastly, semantic consistency

¹⁰ ISO/IEC 25024:2015, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. Available: <https://www.iso.org/standard/35749.html>

is maintained by verifying that data values conform to predefined semantic rules, ensuring their correctness in the context of AI-based automotive IDS applications.

3.4 Diversity

Diversity refers to the variety among samples in a dataset, ensuring that the data sufficiently represents different conditions, categories, and attack scenarios. Maintaining diversity is crucial in automotive IDS to prevent model bias and ensure generalizability across various real-world security threats. Diversity is assessed through multiple key aspects, including label richness, relative label abundance, and category size diversity. Label richness measures the number of distinct labels present in a dataset, ensuring that different attack types and normal traffic conditions are well represented. Relative label abundance evaluates the distribution of labels, ensuring that no single category dominates the dataset, which could lead to biased model performance. Category size diversity quantifies the balance of categorized samples, identifying whether certain attack or normal traffic classes are underrepresented.

Table 7. SOME/IP diversity.

SOME/IP Diversity	Score	Explanation
Label richness	5 different labels	4 attacks and 1 non-attack label
Relative label abundance	5.27% attack and 94.73% non-attack	See Figure 4
Category size diversity	50%	50% of categories under 50% threshold in each file

Table 8. SAD diversity.

SAD Diversity	Score	Explanation
Label richness	4 different labels	3 attacks and 1 non-attack label from Sonata, Kia, and Spark
Relative label abundance	10.56% attack and 89.44% non-attack	See Figure 5
Category size diversity	50%	50% of categories under 50% threshold in each file

Table 7 and Table 8 show the results of SOME/IP and SAD diversity. There are 5 labels (4 attacks and 1 non-attack lable) in SOME/IP and 4 labels (3 attacks and 1 non-attack label taken from Sonata, Kia and Spark vehicles). The label distribution between the two classes (i.e., 0 representing normal and 1 representing attack) is quite imbalanced for both SOME/IP and the SAD dataset as can also be seen in Fig 4 and Fig. 5 for each attack types. Also, the category size diversity for each category is 50% and is lower than a 50% threshold in each file.

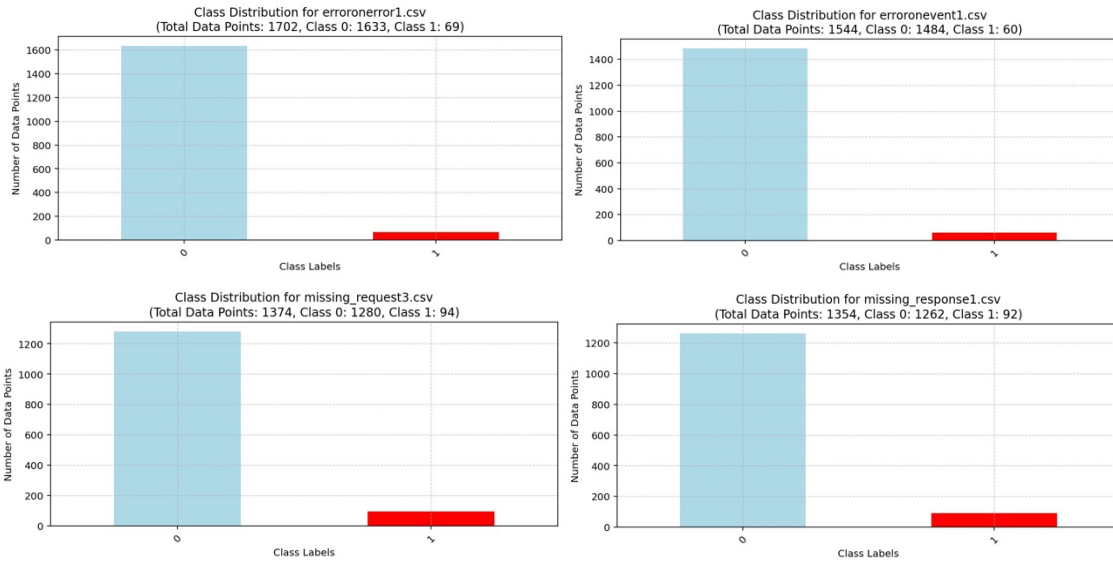
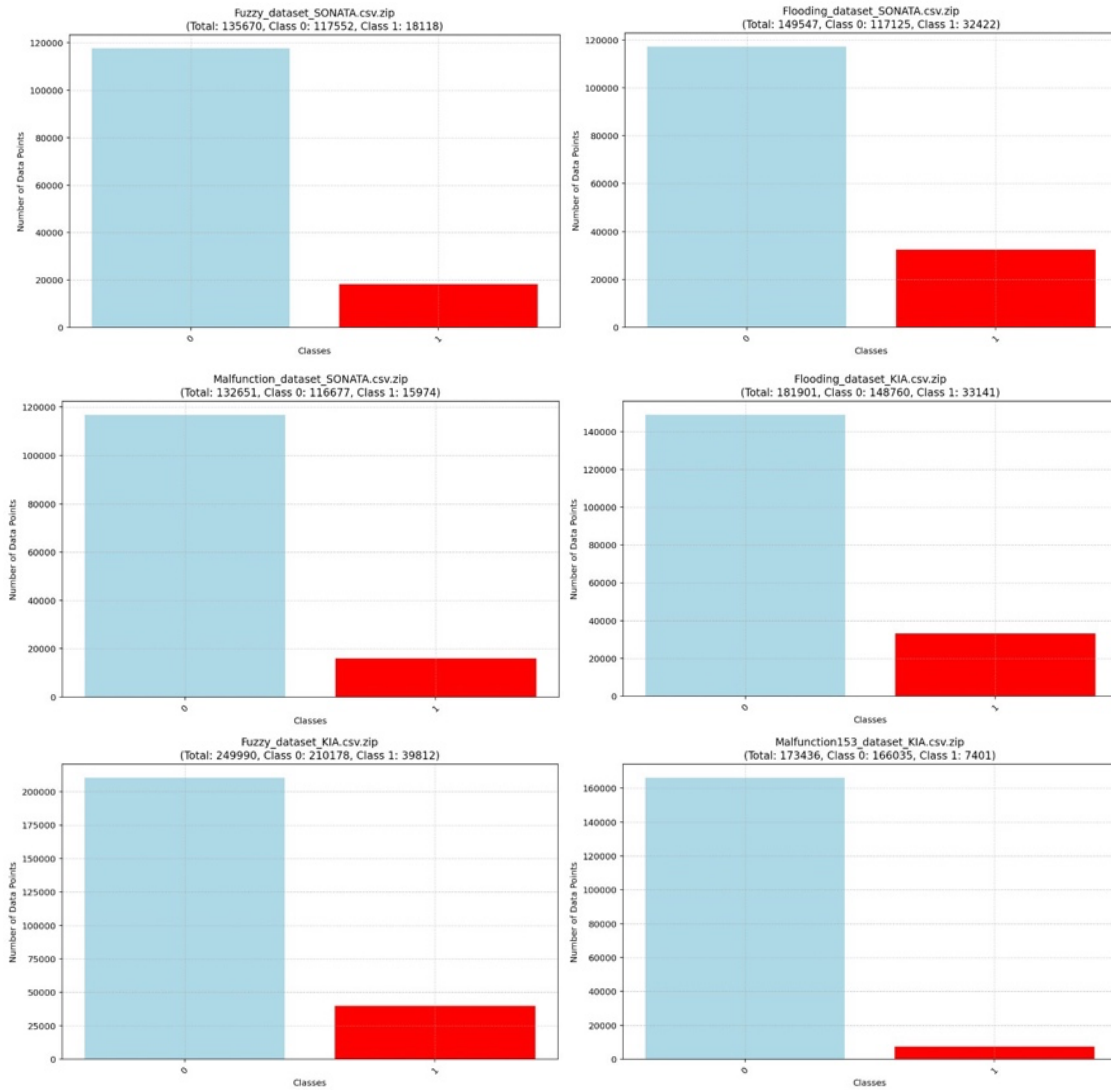


Fig. 4 SOME/IP label distribution.



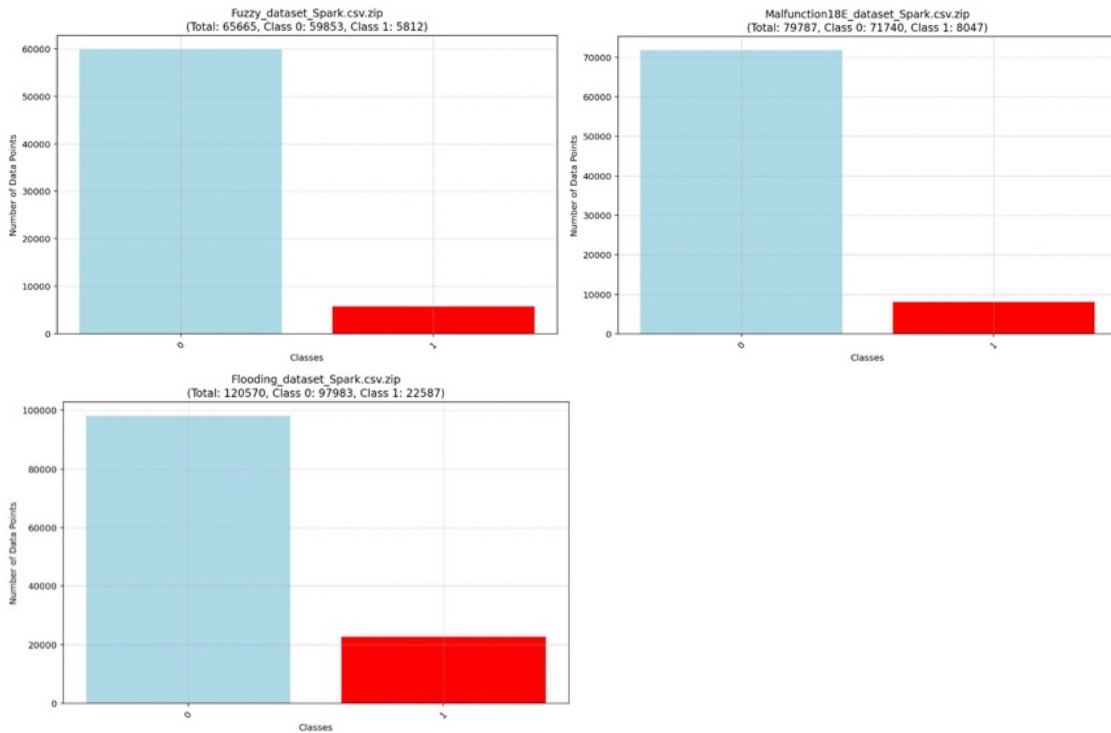


Fig. 5 SAD label distribution.

Ensuring adequate class representation is critical, particularly in attack vs. normal classifications, or more granular evaluations such as specific attack types (e.g., Spoofing vs. DoS). The dataset must include sufficiently distinct classes to enable effective IDS model training while preventing an overfitting bias toward more frequent labels.

3.5 Credibility

Credibility, defined as the extent to which data attributes are considered believable and reliable within a specific usage context, plays a critical role in ensuring the trustworthiness of datasets used in intrusion detection systems (IDS). It is evaluated based on value credibility and source credibility, as outlined in ISO/IEC 25024. Value credibility assesses whether the dataset follows expected behavioral patterns and domain expertise. The SOME/IP dataset is developed based on synthetic data reflecting AUTOSAR standard behavior. Similarly, the SAD dataset is derived from real vehicle operating data collected through an in-vehicle OBD-II port, ensuring an accurate representation of network communication in automotive Ethernet systems.

The SOME/IP dataset is sourced from Paris Telecom¹¹, a well-established institution recognized for its contributions to automotive cybersecurity research. For credibility, dataset documentation is important as it provides detailed insights into data origin, attack scenarios, collection methodologies, and compliance with standards. Source credibility ensures that datasets originate from qualified and reputed organizations. The SAD analysis dataset is developed by the Hacking and Countermeasure Research Lab

¹¹ N. Alkhatib, SOMEIP_IDS: Intrusion Detection System for SOME/IP Protocol, GitHub repository, 2023. Available: https://github.com/Alkhatibnatasha/SOMEIP_IDS/tree/main

(HCRL)¹², School of Cybersecurity, Korea University, led by Professor Huy Kang Kim, ranked among Stanford/Elsevier's top 2% scientists in 2024, reinforcing its credibility for both academic and industrial applications. The SOME/IP dataset provides details on simulation procedures, attack types, and adherence to AUTOSAR specifications. The SAD dataset includes comprehensive documentation on CAN-bus data origins and attack methodologies. Additionally, clear objective and parseability as demonstrated by both datasets improves credibility by ensure that dataset structures align with their intended security applications and applicable for automotive IDS research. Together, these factors validate the robustness, reliability, and practical relevance of the datasets for evaluating and improving intrusion detection mechanisms in automotive networks.

3.6 Currentness

Currentness refers to the time difference (ΔT) between when a data sample is recorded and when it is used, ensuring that datasets remain relevant to evolving technological standards and emerging cybersecurity threats. This characteristic is crucial in evaluating whether datasets align with modern security challenges and system behaviors. Currentness is assessed based on feature currentness and record currentness, with a 5 and 15-year threshold serving as a benchmark for determining a dataset's applicability to contemporary threat landscapes.

Feature currentness evaluates whether individual data features fall within the required recency window. And record currentness assesses whether all data records within a dataset remain within the required time window. The SAD dataset, created in 2018, remains within a 15-year threshold but surpasses a 5-year threshold when assessed in 2025 for feature and record currentness, suggesting that it might not fully capture up-to-date attack behaviors or vulnerabilities of present-day. Recent advancements in automotive technology have introduced new vulnerabilities in CAN bus systems. For instance, in 2022, a car theft incident demonstrated how thieves accessed the CAN bus via the headlight wiring to start the engine without a key¹³. In contrast, the SOME/IP dataset, created in 2023, remains within a 5-year threshold, ensuring its features and records align with current technological standards and the latest developments in automotive Ethernet-based intrusion detection systems (IDS). Both the datasets are within a 15-year threshold assuming slow industry change.

Overall, the SOME/IP dataset exhibits higher currentness than the SAD dataset, as it reflects more recent network-based intrusion detection challenges. However, the SAD dataset retains its value for historical analysis, providing insights into long-term trends in CAN-bus attack patterns and IDS effectiveness over time. While the SOME/IP dataset is more relevant for modern network security applications, the SAD dataset continues to serve as a valuable resource for studying legacy systems and historical attack behaviors.

¹² H. Kim, J. Woo, and H. Kim, "SURVIVAL: Real-Time Intrusion Dataset for In-Vehicle Networks," Korea University Cyber Security Lab. Available: <https://ocslab.hksecurity.net/Datasets/survival-ids>

¹³ O. Yang, "How to Get Away With Car Theft: Unveiling the Dark Side of the CAN Bus," VicOne, May 5, 2023. [Online]. Available: <https://vicone.com/blog/how-to-get-away-with-car-theft-unveiling-the-dark-side-of-the-can-bus>

4 Observations and conclusion

The evaluation of the SAD and SOME/IP datasets highlights key areas of strength and improvement in terms of data quality for automotive intrusion detection. Accuracy and completeness assessments indicate that both datasets meet fundamental quality criteria for anomaly detection; however, expanding the range of attack scenarios in the SOME/IP and SAD dataset would improve its completeness. Diversity analysis suggests that both the SOME/IP and SAD dataset could be more balanced, if certain attack types were not underrepresented, ensuring a fair class distribution for model training. Regarding credibility, extended dataset documentation—particularly for SOME/IP—would improve trust and usability. Lastly, currentness remains a critical factor; updating the SAD dataset with more recent CAN-bus data and incorporating modern attack patterns would ensure continued relevance in evolving cybersecurity landscapes. The SAD and SOME/IP datasets are highly valuable for autonomous vehicle research and intrusion detection. However, some gaps, particularly in attack diversity and bias in data distribution, should be addressed to fully meet the AI Act requirements. The alignment of the datasets with ISO/IEC data quality standards enhanced the understandability, reliability and overall trustworthiness for automotive cybersecurity applications.

Through our international collaboration programmes with academia, industry, and the public sector, we ensure the competitiveness of the Swedish business community on an international level and contribute to a sustainable society. Our 2,800 employees support and promote all manner of innovative processes, and our roughly 100 testbeds and demonstration facilities are instrumental in developing the future-proofing of products, technologies, and services. RISE Research Institutes of Sweden is fully owned by the Swedish state.

I internationell samverkan med akademi, näringsliv och offentlig sektor bidrar vi till ett konkurrenskraftigt näringsliv och ett hållbart samhälle. RISE 2 800 medarbetare driver och stöder alla typer av innovationsprocesser. Vi erbjuder ett 100-tal test- och demonstrationsmiljöer för framtidssäkra produkter, tekniker och tjänster. RISE Research Institutes of Sweden ägs av svenska staten.



RISE Research Institutes of Sweden AB
Box 857, 501 15 BORÅS, SWEDEN
Telephone: +46 10-516 50 00
E-mail: info@ri.se, Internet: www.ri.se

Applied digitalization
RISE Report : 2025:43
ISBN: