

# X-CANIDS: Signal-Aware Explainable Intrusion Detection System for Controller Area Network-Based In-Vehicle Network

Seonghoon Jeong, *Member, IEEE*, Sangho Lee, Hwejae Lee, and Huy Kang Kim, *Member, IEEE*

**Abstract**—Controller Area Network (CAN) is an essential networking protocol that connects multiple electronic control units (ECUs) in a vehicle. However, CAN-based in-vehicle networks (IVNs) face security risks owing to the CAN mechanisms. An adversary can sabotage a vehicle by leveraging the security risks if they can access the CAN bus. Thus, recent actions and cybersecurity regulations (e.g., UNR 155) require carmakers to implement intrusion detection systems (IDSs) in their vehicles. The IDS should detect cyberattacks and provide additional information to analyze conducted attacks. Although many IDSs have been proposed, considerations regarding their feasibility and explainability remain lacking. This study proposes X-CANIDS, which is a novel IDS for CAN-based IVNs. X-CANIDS dissects the payloads in CAN messages into human-understandable signals using a CAN database. The signals improve the intrusion detection performance compared with the use of bit representations of raw payloads. These signals also enable an understanding of which signal or ECU is under attack. X-CANIDS can detect zero-day attacks because it does not require any labeled dataset in the training phase. We confirmed the feasibility of the proposed method through a benchmark test on an automotive-grade embedded device with a GPU. The results of this work will be valuable to carmakers and researchers considering the installation of in-vehicle IDSs for their vehicles.

**Index Terms**—CAN database, explainability, in-vehicle intrusion detection, self-supervised anomaly detection, UN Regulation No. 155 (UNR 155)

## I. INTRODUCTION

In-vehicle networks (IVNs) are essential for vehicles that are operated by several electronic control units (ECUs). Among the various networking protocols that have been designed for vehicles, Controller Area Network (CAN), which replaces the mesh-like wiring harness with a bus topology, is the

most successful. CAN 2.0A, which was released in 1991, is currently employed in almost all vehicles because it meets the crucial requirements of IVNs. In particular, the bus topology, arbitration mechanism, and short frame enable broadcasting, interconnect multiple ECUs, and prevent medium occupation, respectively. However, these mechanisms are the root cause of security risks in CAN-based IVNs, which allow an adversary to eavesdrop on in-vehicle communication, inject arbitrary messages, and cause denial of service of a specific ECU [1] or an entire CAN bus [2].

Until the early 2010s, adversaries were considered negligible because they must have physical access to a CAN-based IVN to leverage the security risks. However, this situation has changed with the widespread use of connected vehicles, as the connectivity broadens the remote attack surfaces of connected vehicles [3], [4]. Compromised in-vehicle infotainment systems may allow the adversaries to obtain access to IVNs remotely. Previous studies supported this concern [5], [6], [7], [8], [9]. In particular, Miller and Valasek [3] jumpstarted cybersecurity studies on vehicles with the proof of concepts of remote hacking into a CAN-based IVN of a Jeep Cherokee.

It is crucial to protect vehicles from cyberattacks to safeguard passengers and pedestrians against unexpected vehicle behavior. Nevertheless, it is impossible to remedy the security risks of CAN-based IVNs without revising the protocol specifications. Intrusion detection systems (IDSs) have been proposed to identify anomalies in CAN-based IVNs [4], [10]. In the early research stages, statistical approaches and conventional machine-learning algorithms were considered. As deep-learning techniques have evolved, recent studies have tended to adapt such techniques for precise intrusion detection. IDSs are currently becoming an essential component of vehicles. For example, United Nations Regulation No. 155 (UNR 155), which is a recent automotive cybersecurity regulation that will take effect in many countries from 2024, states that “the vehicle manufacturer shall implement measures for the vehicle type to (a) detect and prevent cyber-attacks against vehicles of the vehicle type; (b) support the monitoring capability of the vehicle manufacturer with regards to detecting threats; (c) provide data forensic capability to enable an analysis of attempted or successful cyber-attacks” (see §7.3.5 in [11]).

In this study, we propose a practical IDS named X-CANIDS to address the following three limitations with respect to IDSs for CAN-based IVNs. First, previous IDSs did not provide additional information for forensics. Most carmakers have their own pattern database to distinguish a malfunction. An explanation of detection result can help carmakers analyze

Manuscript received Jan. 10, 2023; revised Jun. 7, 2023, Aug. 30, 2023, and October 18, 2023; accepted October 21, 2023. This research was supported by the 2021 autonomous driving development innovation project of the Ministry of Science and ICT, “Development of technology for security and ultra-high-speed integrity of the next-generation internal network of autonomous vehicles” (No. 2021-0-01348). (Corresponding author: Huy Kang Kim.)

Seonghoon Jeong is with the Institute of Cybersecurity and Privacy, Korea University, Seoul 02841, Republic of Korea (e-mail: seonghoon@korea.ac.kr).

Sangho Lee is with the Samsung Research, Seoul 06765, Republic of Korea (e-mail: s35.lee@samsung.com).

Hwejae Lee and Huy Kang Kim are with the School of Cybersecurity, Korea University, Seoul 02841, Republic of Korea (e-mail: {hwejae94, cenda}@korea.ac.kr).

This is the Accepted version of an article for publication in IEEE TVT. ©2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. Digital Object Identifier 10.1109/TVT.2023.3327275

conducted attacks and prepare a remedy such as updating their database. Supervised methods (e.g., [12], [13]) have been designed to distinguish the type of attack. However, they require ground-truth labels and only distinguish predefined attack types. Second, the evaluations have lacked a feasibility perspective. Evaluation is necessary because an IDS works on an ECU or an in-vehicle component. Regardless of the effectiveness of an IDS, it may be useless owing to bottlenecks. Finally, limited studies have considered the use of signals rather than raw payloads. Signals help to improve the detection performance of an IDS because they reflect the context of the vehicle. Nevertheless, the use of signals has rarely been discussed because of a lack of knowledge regarding payload deserialization methods. To date, two studies have used sensor values through the on-board diagnostics (OBD)-II feature [14] and several reverse-engineered signals [15].

The contributions of this study are summarized as follows:

**1. Self-supervised intrusion detection with signals.** We propose X-CANIDS, which is a novel method that consists of a feature generator and intrusion detection model. The feature generator builds a time-series representation of signals that are deserialized from the payloads of the CAN messages. We use a CAN database to train X-CANIDS with 107 signals. The detection model is trained using an attack-free dataset. X-CANIDS can detect zero-day attacks, of which we are unaware at the time of implementation.

**2. Explainability.** X-CANIDS provides additional information on which systems (i.e., ECUs) and what data were compromised, even if no ground-truth labels or intrusion datasets are available in the training phase. The explainability is beneficial for carmakers and incident response teams to analyze conducted attacks. To this end, we leverage the signalwise reconstruction error combined with an autoencoder.

**3. Feasibility.** Our method is benchmarked on our NVIDIA Jetson AGX Xavier, which is an automotive-grade embedded device. The benchmark results confirm that X-CANIDS is promising for real-world use cases. X-CANIDS achieves a deterministic detection latency of 38.2512–73.2512 ms on the embedded device with a feature generation frequency and batch size of 200 Hz and 8, respectively.

The remainder of this paper is organized as follows. In §II, we provide the background for this study. In §III, we describe the proposed X-CANIDS. §IV outlines the CAN datasets that are used in the experiment. The experimental results are presented and the detection performance, explainability, and feasibility are discussed in §V. In §VI, we categorize prior studies based on the features that are used to detect in-vehicle intrusion. Finally, we conclude the study in §VII.

## II. PRELIMINARIES

### A. Terminology

We present the terminology used in this paper. The *payload* is a bit sequence that is encapsulated in the data field of a CAN frame (see Fig. 1). The *signal* indicates a sensor value that is deserialized from a portion of the payload. A *stream* is a set of CAN messages with a particular arbitration ID. The aim of this study is to develop an *explainable* and a *feasible* IDS. We

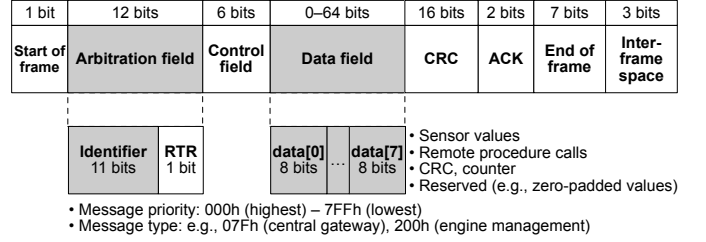


Fig. 1. CAN 2.0A frame structure. An ECU application refers to the arbitration and data field.

can state that the proposed IDS is explainable when a detection result allows us to understand which signal or ECU is affected by attacks. We can state that the proposed IDS is feasible when it can evaluate all given inputs within a deterministic latency on an ECU or an automotive-grade embedded device.

### B. CAN Frame

The CAN frame is a communication unit between two in-vehicle ECUs that are connected by a CAN bus. Fig. 1 depicts the CAN 2.0A frame structure. Among the fields, the ECU firmware uses the 11-bit arbitration identifier (AID) field and a 64-bit data field. The AID is considered as a categorical value for determining the message type as well as a numerical value that represents the priority of the message. The data field contains a payload that represents various signals that are used to operate vehicles, such as sensor values, Booleans, inter-ECU procedure calls, cyclic redundancy check values, sequential counters, and zero padding. As a CAN frame does not contain a transmission timestamp, the CAN receiver assigns a timestamp for each inbound CAN message using its internal clock. In the ECU firmware, a transmitted CAN frame is represented as a CAN message  $m \rightarrow (t, a, \mathbf{p})$ , where the timestamp  $t \geq 0$ , AID  $a \in \{0, 1, \dots, 2047\}$ , and bit sequence vector of the payload  $\mathbf{p} = \{p_i | p_i \in \{0, 1\} \text{ for } i = 1..n\}$ , with  $n \in \{0, 8, 16, \dots, 64\}$ .

The ECU firmware serializes one or more signals in the data field prior to transmission. Each receiver can deserialize the payload of the data field to use the original values. Researchers can easily obtain a CAN dataset  $\mathcal{M} = \{m_1, m_2, m_3, \dots\}$  through their CAN nodes by leveraging the broadcast nature of the CAN. However, it is difficult for them to understand the specific meaning of  $m$  owing to the lack of information regarding the exact representation of the given  $a$  and  $\mathbf{p}$ . Meanwhile, researchers who wish to build after-market autonomous driving kits or use in-vehicle signals for specific purposes, such as building a payload-based IDS for CAN buses [15], are motivated to reverse engineer CAN message payloads. Because manual reverse engineering requires substantial effort, several automated methods have been proposed for this purpose [16], [17], [18], [19], [20], [21], [22], and increasingly sophisticated results have been achieved in recent years. However, these methods remain insufficient for deserializing hundreds of signals precisely.

B0_902_WHL_SPD11: 8 ABS									
SG_WHL_SPD_FL	:	0 14@1+	(0.03125, 0.0)	[0.0 511.96875]	"km/h"	-4WD, AFLS, ...			
SG_WHL_SPD_FR	:	16 14@1+	(0.03125, 0.0)	[0.0 511.96875]	"km/h"	-4WD, ACU, ...			
SG_WHL_SPD_RL	:	32 14@1+	(0.03125, 0.0)	[0.0 511.96875]	"km/h"	-4WD, AFLS, ...			
SG_WHL_SPD_RR	:	48 14@1+	(0.03125, 0.0)	[0.0 511.96875]	"km/h"	-4WD, AFLS, ...			
SG_WHL_SPD_AliveCounter_LSB	:	14 2@1+	(1.0, 0.0)	[0.0 3.0]	""	-4WD, EMS, LPI, ...			
SG_WHL_SPD_AliveCounter_MSB	:	30 2@1+	(1.0, 0.0)	[0.0 3.0]	""	-4WD, EMS, LPI, ...			
SG_WHL_SPD_Checksum_LSB	:	46 2@1+	(1.0, 0.0)	[0.0 3.0]	""	-4WD, EMS, LPI, TCU, TMU			
SG_WHL_SPD_Checksum_MSB	:	62 2@1+	(1.0, 0.0)	[0.0 3.0]	""	-4WD, EMS, LPI, TCU, TMU			

Signal name	Bit index   bit length	(scale, offset)	[min   max]	unit	Other ECUs using the signal
	Endianness (0 big, 1 little)				
	Signness (+ unsigned, - signed)				

Fig. 2. Snippet of CAN database *hyundai\_2015\_ccan.dbc* [23].**Algorithm 1:** Deserialization of a CAN message.

---

**Input:**  $a, p$ : AID and payload of CAN message  
**Data:** DBC: CAN database  
**Output:**  $s = \{s_i | i = 1..n\}$ : List of deserialized signals

```

1 signals  $\leftarrow$  get_signal_specification(DBC,  $a$ )
2  $n \leftarrow |\text{signals}|$  // Number of signals
3 Initialize a new vector  $s \in \mathbb{R}^n$ 
4 for  $i \leftarrow 1$  to  $n$  do
    // Parse the specification of a signal
5     bit_idx, bit_len, endianness, signness, scale, offset, min,
       max  $\leftarrow$  signal $_i$ 
    // Obtain a subset of bit sequence
6      $p' \leftarrow \{p_{\text{bit\_idx}}, \dots, p_{\text{bit\_idx}+\text{bit\_len}}\}$ 
    // Decode the bit sequence into an integer
7      $s_i \leftarrow \text{int}(p', \text{endianness}, \text{signness})$ 
    // Scale the signal
8      $s_i \leftarrow s_i / \text{scale} + \text{offset}$ 
9     if not  $\min \leq s_i \leq \max$  then
10        Raise an error.
```

---

**C. CAN Database**

The CAN database is a network dissector that consists of formal payload deserialization descriptions. We introduce the CAN database using a straightforward example, because signals that are deserialized from CAN messages are crucial inputs for the proposed method. A CAN database describes the specification of a certain CAN-based IVN, including the bitrate, list of ECUs, signals, and Tx-Rx ECU relationships. The CAN database specifies the bit indices, endianness, signness, scale, offset, value range, units, and list of ECUs that refer to each signal. In general, CAN databases are composed by carmakers at the time of the IVN design.

Recently, Comma.ai, which is a company that develops after-market autonomous driving kits, released CAN databases that work with some commercialized vehicles in a public Git repository known as OpenDBC [23]. The CAN databases are formatted in the well-known DBC file format that was introduced by Vector Informatik GmbH. Fig. 2 presents a snippet of a CAN database that was obtained from the repository. The first line states that an ECU named ABS (*i.e.*, anti-lock braking system) transmits the message WHL\_SPD\_11, which consists of  $a = 902$  and  $|p| = 64$  (*i.e.*, 8 bytes). The remainder defines eight signals of rotation speeds (front left, front right, rear left, and rear right wheels) that are represented in km/h units, two checksum values, and two alive counters. The meanings of the CAN database syntax are annotated at the bottom of the figure.

The deserialization procedure  $D$  is described in Algo-

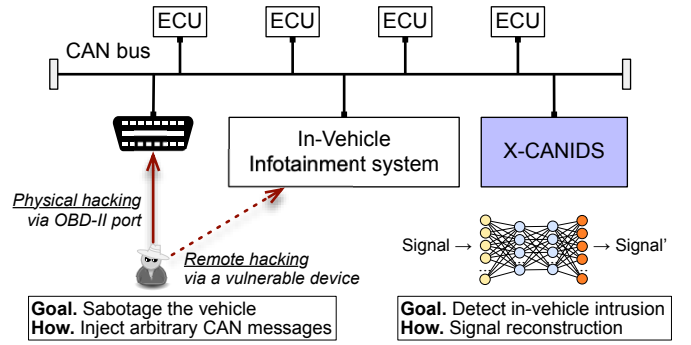


Fig. 3. Considered in-vehicle network architecture.

**Algorithm 1.** The deserialization procedure can be represented by  $D(a, p) = s = \{s_i | s_i \in \mathbb{R} \text{ for } i = 1..n\}$ , where  $n$  is the number of signals that are defined in the given DBC (see lines 1–2). The output vector  $s$  contains human-understandable signals. The loop (lines 4–10) can be executed concurrently because  $s_i$  are independent of one another.

**D. Adversary and Attack Model**

In this study, we consider an adversary who wants to sabotage a vehicle by injecting arbitrary CAN messages. Fig. 3 depicts the supposed adversary in the CAN-based IVN. The adversary must obtain access to the target CAN-based IVN to conduct an attack. The adversary may consider physical hacking by installing a CAN dongle at the OBD-II port. The attacker may also consider the remote exploitation of vehicle-to-everything communication-enabled ECUs, such as an infotainment system [6], [7], [8], [9]. Once the adversary obtains access, they can conduct five types of attacks [1], [15], as follows:

1) *Fuzzing Attack*: It manipulates various ECUs with random payloads and it can be performed with CAN messages that contain random AIDs and payloads. The attack can cause a malfunction of the target vehicle even if the adversary does not have prior knowledge of the in-vehicle communications.

2) *Fabrication Attack*: A specific ECU is manipulated with the intention of the adversary, and it can be performed using well-crafted CAN messages with a specific AID and payload. As a legitimate ECU periodically transmits CAN messages with the same AID, an adversary can transmit their CAN message directly after every benign message.

3) *Suspension Attack*: It neutralizes an ECU by exploiting the error-handling mechanism of the CAN [1]. A target ECU does not transmit any CAN messages during the attack.

4) *Masquerade Attack*: It is a combination of the fabrication and suspension attacks. A stream from a specific ECU is replaced with arbitrary messages that are generated by the adversary during the attack.

5) *Replay Attack*: An adversary captures legitimate CAN messages in a certain period. Then, they transmit the CAN messages within the CAN bus. The attack can cause a certain malfunction that the target vehicle have performed in the capture duration.

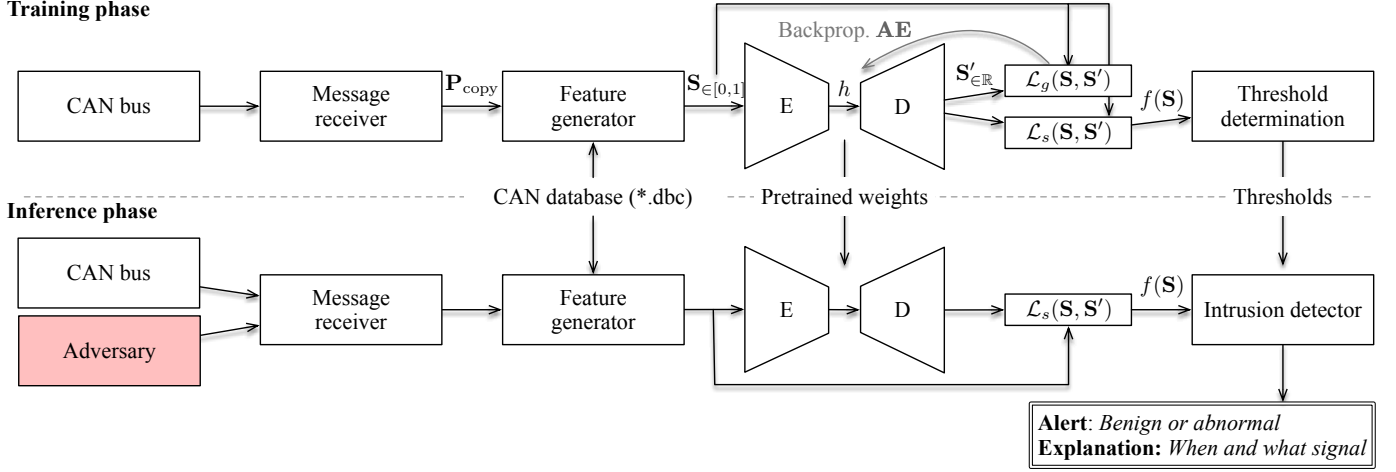
**Training phase**

Fig. 4. Proposed framework. In the training phase, the proposed framework uses attack-free CAN messages to train the autoencoder and to determine the threshold. In the inference phase, the proposed framework determines the signal that is affected by the adversary if the CAN bus is under attack.

**III. METHODOLOGY**

In this section, we present X-CANIDS, which consists of a message receiver, a feature generator, an autoencoder, and a decision-maker. As illustrated in Fig. 3, the proposed in-vehicle IDS is directly connected to the CAN-based IVN to receive all CAN messages instantly. It is a preferable architecture for prior CAN IDSs, while maintaining a simple topological structure. Fig. 4 outlines the proposed framework. In the training phase, X-CANIDS uses benign CAN messages to train the autoencoder AE towards small global errors. At the end of the training phase, X-CANIDS determines the thresholds using signalwise errors. In the inference phase, X-CANIDS determines whether a given input is affected by an adversary using pretrained weights and thresholds. Signalwise errors are considered to explain the detected attack. The CAN database is required in both phases for the feature generator to deserialize the signals from the CAN messages.

**A. Message Receiver**

The message receiver is connected to the CAN bus to monitor all CAN messages. The message receiver contains the matrix  $\mathbf{P} \in \{\emptyset, 0, 1\}^{N \times M}$  to cache the latest payload of each stream, where  $N$  is the number of streams in the CAN bus and  $M$  is the maximum length of the payload (64 bits for the CAN 2.0A bus and 512 bits for the CAN-FD bus).

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NM} \end{pmatrix} \quad (1)$$

In (1), the  $n$ -th row of the matrix  $\mathbf{P}$  represents the latest payload of a certain stream for each  $\mathbf{p}_n$ . Initially,  $p_{n,m} = \emptyset \forall n \in \{1..N\}, m \in \{1..M\}$ . As the message receiver monitors each CAN message from the CAN bus,  $p_{n,m}$  becomes 0 or 1. It should be noted that  $\mathbf{P}$  is volatile because the elements change continuously upon the arrival of CAN messages.

**B. Feature Generator**

The feature generator interprets each  $\mathbf{P}$  into a feature matrix  $\mathbf{S}$ , which is fed to the autoencoder. The feature generator pipeline consists of the following: (1) the payload sampler, (2) deserializer, (3) feature scaler, and (4) time-series feature generator.

1) *Payload Sampler*: The payload sampler captures  $\mathbf{P}_{copy}$ , which is a static copy of  $\mathbf{P}$ , from the message receiver in every time interval  $t$ . The payload sampler begins working when  $\mathbf{P}$  satisfies  $p_{n,1} \neq \emptyset \forall n \in \{1..N\}$ , which means that the message receiver observes each stream at least once. Each copy is forwarded to the deserializer.

2) *Deserializer*: The deserializer converts a given  $\mathbf{P}_{copy}$  into a vector  $\mathbf{s} \in \mathbb{R}$  that contains human-understandable signals, such as the engine speed and steering wheel angle. The deserialization procedure for  $\mathbf{P}_{copy}$  can be represented as

$$\mathbf{P}_{copy} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \end{pmatrix} \xrightarrow{\begin{matrix} D(a_1, \mathbf{p}_1) \\ D(a_2, \mathbf{p}_2) \\ \vdots \\ D(a_N, \mathbf{p}_N) \end{matrix}} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{pmatrix} \xrightarrow{\text{concatenate}} \mathbf{s}. \quad (2)$$

The streamwise deserialization procedure calculates Algorithm 1. Thus, the output vectors  $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N$  are dependent on the given CAN database and  $\mathbf{P}_{copy}$ . In particular, the CAN database determines the number of elements in  $\mathbf{s}_i$ . The vector  $\mathbf{s} = \parallel_{n=1}^N \mathbf{s}_n$  represents the concatenation of all deserialized signals and is fed to the feature scaler.

3) *Feature Scaler*: The feature scaler is crucial because the elements of  $\mathbf{s}$  have various value ranges. For example, the engine speed and steering wheel angle can be represented within  $[0, 8191]$  RPM and  $[-1024, 1023]$  degrees, respectively. Furthermore, a binary value (0 or 1) can be used to represent whether a foot brake is engaged. The variation in signal ranges may result in unstable training of AE because it can be considered as a weight (i.e., feature importance).

We design a lightweight feature scaler to achieve robust performance. The signal scaler is designed to normalize a

given  $\mathbf{s}$  to  $\hat{\mathbf{s}} = \{\hat{s}_i | \hat{s}_i \in [0, 1], i = 1..x\}$ , where  $x = |\mathbf{s}|$  is the number of concatenated signals. The proposed feature scaling procedure is presented in (3), where the parameters  $\min_i$  and  $\max_i$  are the minimum and maximum values, respectively, of the  $i$ -th signal described in the CAN database.

$$\hat{s}_i = \frac{s_i - \min_i}{\max_i - \min_i} \quad (3)$$

Our feature scaler is the same as the conventional min-max scaler, except for the determination method of the parameters  $\min_i$  and  $\max_i$ . If we had applied the min-max scaler, it would fit the parameters according to observations in the training phase (*i.e.*, a training set). A downside of the min-max scaler is that it does not correctly handle outliers that are over the maximum or under the minimum values that may be observed in the inference phase. We revised the min-max scaler to this end. Our feature scaler leverages the minimum and maximum values that are predefined in the CAN database to overcome this drawback. Note that a scaled signal  $\hat{s}_i$  from our feature scaler is always represented by 0–1 because  $s_i$  satisfies  $\min_i \leq s_i \leq \max_i$  (cf. lines 9–10 of Algorithm 1). The scaled vector  $\hat{\mathbf{s}}$  is fed to the time-series feature generator.

4) *Time-Series Feature Generator*: The time-series feature generator builds an input for the autoencoder in a sliding-window manner. It temporarily remembers the most recent  $w$  input vectors and returns a two-dimensional matrix  $\mathbf{S} \in [0, 1]^{w \times x}$  by stacking them. The window size  $w$  is a parameter that determines the number of time steps that are contained by the feature. The time-series feature generator returns  $\mathbf{S}$  every  $t$  as the payload sampler feeds  $\mathbf{P}_{\text{copy}}$ . The matrix representation helps the autoencoder to understand the time series and lateral relationships between the signals.

### C. Autoencoder

In the proposed framework, the autoencoder  $\mathbf{AE}$  is adapted to model the attack-free state of a moving vehicle. The autoencoder is a self-supervised neural network that consists of an encoder and a decoder, as expressed by

$$\mathbf{AE}(\mathbf{S}) = \text{Decoder}(\text{Encoder}(\mathbf{S})) = \text{Decoder}(h) = \mathbf{S}' \quad (4)$$

The encoder compresses the input feature  $\mathbf{S}$  into a low-dimensional latent vector  $h$ . Subsequently, the decoder attempts to reconstruct the original input data as far as possible using the latent vector. In the training phase,  $\mathbf{AE}$  is fitted with features from benign datasets using backpropagation to reduce the global mean squared error (MSE). The global MSE is calculated as follows:

$$\mathcal{L}_g(\mathbf{S}, \mathbf{S}') = \frac{1}{wx} \sum_{j=1}^w \sum_{i=1}^x (\mathbf{S}_{ji} - \mathbf{S}'_{ji})^2. \quad (5)$$

It is also referred to as the reconstruction error in terms of an autoencoder. The goal of  $\mathbf{AE}$  is to exhibit a small reconstruction error in the inference phase, particularly when a given sample is attack free. However,  $\mathbf{AE}$  is required to exhibit a high reconstruction error when the input  $\mathbf{S}$  is affected by an adversary.

$\mathbf{AE}$  is supposed to be computed on an automotive-grade embedded device. Therefore, the model complexity should be considered while minimizing the reconstruction error. We conceive the six candidate layers for  $\mathbf{AE}$  as follows: the fully connected, 1D and 2D convolutional, 1D separable convolutional, long short-term memory (LSTM), and bidirectional LSTM (BiLSTM) layers. We evaluate the layers using training and validation datasets.

### D. Intrusion Detection and Explanation

The global MSE can be used solely to distinguish anomalies. Nevertheless, we measure the signalwise MSE to obtain explainable intrusion detection results. In this section, we define the inference function  $f(\mathbf{S})$  that calculates the signalwise MSE. Thereafter, we introduce the threshold determinator and intrusion detector.

1) *Inference Function*: Once  $\mathbf{AE}$  has been fitted, the signalwise loss function  $\mathcal{L}_s$  is combined with  $\mathbf{AE}$  to formulate the inference function  $f(\mathbf{S})$ . Equation (6) presents the inference function that returns the loss vector  $\mathbf{l}$ . The element  $l_i$  is the loss of the  $i$ -th signal.

$$\begin{aligned} f(\mathbf{S}) &= \mathcal{L}_s(\mathbf{S}, \mathbf{AE}(\mathbf{S})) = \mathcal{L}_s(\mathbf{S}, \mathbf{S}') \\ &= \frac{1}{w} \sum_{j=1}^w (\mathbf{S}_j - \mathbf{S}'_j)^2 = \mathbf{l} = \{l_1, l_2, \dots, l_x\} \end{aligned} \quad (6)$$

2) *Threshold Determination*: The threshold determinator is used during the training phase. This module aims to determine  $\theta_i$  for the error rate calculation of the  $i$ -th signal and the detection threshold  $\Theta$  to raise the alarm.

First, the module calculates a set of loss vectors  $\{\mathbf{l}_1, \mathbf{l}_2, \dots\}$  using the entire training set. Second,  $\theta_i = \bar{l}_i + 3\sigma_i$  is considered for the  $i$ -th signal, where  $\bar{l}_i$  and  $\sigma_i$  are the mean and standard deviation of  $l_i$ s in the set, respectively. Third, the module measures a set of error-rate vectors  $\{\mathbf{r}_1, \mathbf{r}_2, \dots\}$  using the entire validation set. The error rate vector  $\mathbf{r}$  is derived as follows:

$$\mathbf{r} = \{r_i | r_i = l_i / \theta_i \text{ for } i = 1..x\}. \quad (7)$$

Finally,  $\Theta$  is determined by the  $q$ -th percentile of  $\max(\mathbf{r})$  for all  $\mathbf{r}$ s in the set, where  $0.95 \leq q \leq 1$  and  $q$  is a hyperparameter that determines the detection sensitivity.

3) *Intrusion Detection and Explanation*: The intrusion detection module uses the  $\theta_i$ s and  $\Theta$  from the threshold determinator at the beginning of the inference phase. The intrusion detector obtains  $\mathbf{r}$  for each  $\mathbf{S}$  using (6) and (7). The module raises an alarm if  $\mathbf{r}$  satisfies  $\max(\mathbf{r}) > \Theta$ . If an intrusion is identified, the intrusion detector identifies the affected signal index  $i$  using  $\text{argmax}(\mathbf{r})$ .

## IV. DATASETS

The datasets used in the experiment are discussed in this section. Publicly available CAN intrusion datasets (*e.g.*, [2], [24]) exist. However, they could not be utilized because the datasets were stationary or prepared using a dynamometer. Therefore, we captured CAN messages from the Hyundai LF Sonata 2017. We used a Kvaser Memorator Pro 2xHS to

TABLE I  
LIST OF CAN DATASETS.

Dataset	# CAN msg.	Duration	Context	Used for
$\mathcal{M}_1$	3,123,784	23 m 52 s	Driving (18.7 km)	Training
$\mathcal{M}_2$	4,134,495	31 m 35 s	Driving (19.3 km)	Training
$\mathcal{M}_3$	3,233,752	24 m 42 s	Driving (19.4 km)	Training
$\mathcal{M}_4$	4,761,315	36 m 23 s	Driving (21.7 km)	Training
$\mathcal{M}_5$	2,915,969	22 m 17 s	Driving (18.8 km)	Validation
$\mathcal{M}_6$	4,279,902	32 m 42 s	Driving (18.8 km)	Testing
$\mathcal{M}_7$	4,817,589	36 m 54 s	Stationary (0 km)	Training

leverage the high-precision embedded clock while capturing the in-vehicle CAN messages. The device was connected to our vehicle through an OBD-II port, which allowed us to access the chassis–CAN bus. We prepared the seven CAN datasets that are listed in Table I. We captured six datasets while driving a vehicle on urban roads in Seoul, Republic of Korea. One dataset was captured while the vehicle was stationary. The datasets are listed in chronological order. For instance,  $\mathcal{M}_2$  was collected after we have collected  $\mathcal{M}_1$ .

As shown in Fig. 4, X-CANIDS consists of the training and inference phases. In the training phase, pre-captured datasets are used to train an autoencoder. In the inference phase, the trained model processes a live stream. In this paper, we try not to merge entire datasets and conduct an N-fold cross-validation, in which subsets of  $\mathcal{M}_i \forall i \in \{1..6\}$  are used in both training and testing. Instead, we selected old datasets  $\mathcal{M}_1$ – $\mathcal{M}_4$  as the training set, new dataset  $\mathcal{M}_5$  as the validation set, and the newest  $\mathcal{M}_6$  as the test set. We believe such a selection can reflect a real-world use case.

We used the CAN database *hyundai\_2015\_ccan.dbc* that is available on OpenDBC [23] to deserialize the payloads of the datasets into signals. We manually confirmed that the CAN database allowed us to acquire appropriate sensor values, such as the steering wheel angle, temperature, velocity, tire pressure, and door open state.

1) *Overview of Datasets:* The dataset  $\mathcal{M}_1$  is summarized in Table II to introduce our CAN datasets. The chassis–CAN bus consisted of 62 unique streams. We denote the name of the transmitting ECU using the CAN database for each stream. The ECU names; for example, engine management system (EMS), transmission control unit (TCU), and motor-driven power steering system (MDPS), imply that the streams take charge of the communications among critical vehicular applications.

We measured the mean and standard deviation of the message intervals for each stream. ECUs transmit their messages periodically through a well-known mechanism. Thus, it can be observed that the mean time intervals of the streams were approximately one of 0.01, 0.02, 0.05, 0.1, 0.2, 1, and 2 s. However, several streams consisted of nonperiodic messages, namely 044h, 52Ah, 541h, and 553h (where the standard deviation  $\geq 0.01$ ). These streams may not be covered by time-interval-based intrusion detection methods (e.g., [25], [26], [27], [28]) that leverage the periodicity.

The number of signals that were defined in each stream is also summarized. In total, there were 688 types of signals. Furthermore, the number of unique payloads for each stream

TABLE II  
SUMMARY OF  $\mathcal{M}_1$  THAT CONSISTS OF 62 STREAMS.

AID	Sender ECU	Mean $\Delta t$	Std. $\Delta t$	DLC	# signals	# uniq. p
042h	DATC12	1.00	0.000261	8	7	1
043h	DATC13	1.00	0.000261	8	24	1
044h	DATC11	0.96	0.180892	8	6	14
07Fh	CGW5	1.00	0.000124	8	25	1
080h	EMS_DCT11	0.01	0.000304	8	10	105,446
081h	EMS_DCT12	0.01	0.000356	8	6	1,000
111h	TCU11	0.01	0.000191	8	13	555
112h	TCU12	0.01	0.000201	8	12	6,083
113h	TCU13	0.01	0.000203	8	18	441
153h	TCS11	0.01	0.000132	8	29	15
162h	TCU_DCT13	0.01	0.000207	3	3	6,185
164h	VSM11	0.01	0.000220	4	6	16
18Fh	EMS_H12	0.01	0.000449	8	21	2,117
200h	EMS20	0.01	0.000370	6	3	348
220h	ESP12	0.01	0.000211	8	14	130,706
251h	MDPS12	0.01	0.000311	8	11	116,572
260h	EMS16	0.01	0.000664	8	15	41,585
2B0h	SAS11	0.01	0.000683	5	5	17,265
316h	EMS11	0.01	0.000823	8	13	63,371
329h	EMS12	0.01	0.000352	8	19	7,776
381h	MDPS11	0.02	0.000530	8	13	32,040
383h	FATC11	0.02	0.000395	8	19	320
386h	WHL_SPD11	0.02	0.000399	8	8	61,540
387h	WHL_PUL11	0.02	0.000418	6	9	60,269
410h	CGW_USM1	0.20	0.000522	8	17	1
436h	PAS11	0.05	0.000661	4	12	1
47Fh	ESP11	0.02	0.000256	6	11	256
490h	EPB11	0.05	0.000515	7	14	1
492h	EMS19	0.05	0.000460	8	13	4
4F1h	CLU11	0.02	0.000446	4	12	14,476
500h	ACU14	0.10	0.000504	1	3	1
502h	TCU14	0.10	0.000374	4	7	1
507h	TCS15	0.10	0.000288	4	11	1
50Ch	CLU13	0.10	0.000463	8	17	1,484
520h	CGW3	0.10	0.000850	8	4	1
522h	GW_IPM_PE_1	0.20	0.000558	8	10	1
52Ah	CLU15	0.10	0.076209	8	15	94
533h	—	0.10	0.000706	8	—	29
534h	—	0.10	0.000704	8	—	517
535h	—	0.10	0.000674	8	—	1,249
541h	CGW1	0.10	0.014478	8	43	4
544h	—	0.20	0.000422	8	—	1
545h	EMS14	0.10	0.000335	8	8	234
547h	EMS15	0.10	0.000545	8	12	38
549h	BAT11	0.10	0.000570	8	9	6,635
54Ch	TCU_DCT14	0.20	0.000385	8	2	1
553h	CGW2	0.20	0.013671	8	41	6
555h	FPCM11	0.10	0.000507	8	9	670
556h	EngFrzFrm11	0.10	0.000578	8	6	11,649
557h	EngFrzFrm12	0.10	0.000587	8	6	4,932
559h	CGW4	0.20	0.000655	8	23	1
57Fh	HU_MON_PE_01	2.00	0.000782	8	1	1
587h	TMU11	0.20	0.000347	8	8	15
58Bh	LCA11	0.10	0.001104	8	18	7
593h	TPMS11	0.20	0.000554	6	12	11
5A0h	ACU11	1.00	0.001139	8	14	2
5B0h	CLU12	1.00	0.000780	4	1	184
5B4h	—	1.00	0.000776	8	—	1
5BEh	—	1.00	0.000834	8	—	1
5C0h	GW_Warning_PE	1.00	0.000810	8	7	1
5D3h	HU_DATC_PE_00	1.00	0.000806	8	3	1
5FAh	ODS11	1.00	0.000965	8	10	1

is presented. Interestingly, although we captured  $\mathcal{M}_1$  while we drove the vehicle for more than 18 km, certain streams (e.g., 042h, 043h, and 5D3h) had a static payload.

2) *Payload Dynamics:* We expected that the payloads in the CAN messages would change more dynamically when the vehicle moved. To prove this hypothesis, we calculated the bitwise Hamming distance vector  $\mathbf{d}$  for each stream, as follows:

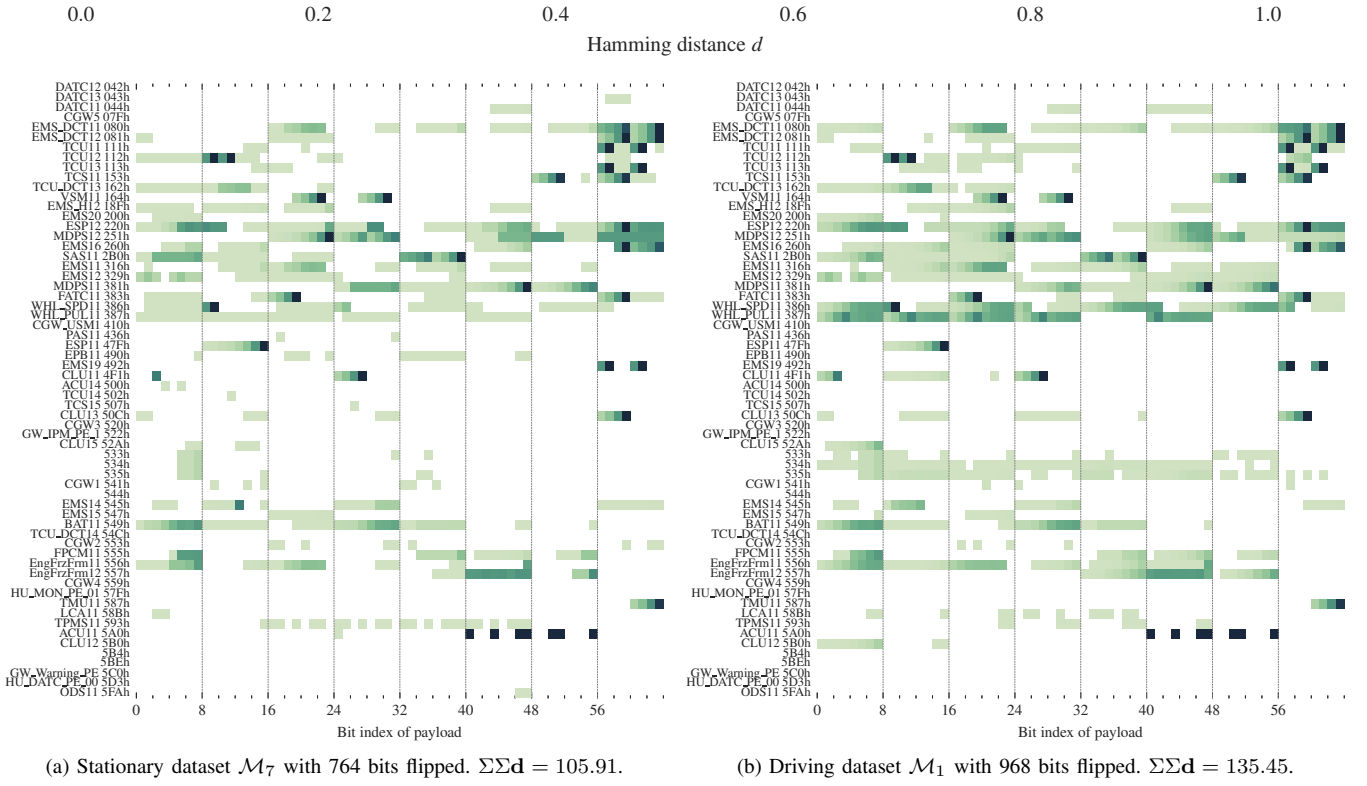


Fig. 5. Bitwise Hamming distance measurements to compare the payload dynamics of two CAN datasets captured during idling and driving. Each cell represents a value  $[0, 1]$  calculated by the number of bits flipped over the observation count. A dark cell ■ means that a bit was flipped nearly every time a message arrived. A light green cell ■ means that a bit was flipped only once or several times. A blank cell ■ indicates no bit flips. A comparison of the two datasets reveals that the payloads changed more dynamically while the vehicle moved.

$$d = \frac{1}{n-1} \sum_{i=2}^n (\mathbf{p}_i \oplus \mathbf{p}_{i-1}). \quad (8)$$

where  $\oplus$  is the bitwise XOR operator,  $\mathbf{p}_i$  is the payload of the  $i$ -th message in a stream, and  $n$  is the number of messages in a stream. For comparison, we measured the Hamming distance using the stationary dataset  $\mathcal{M}_7$  and driving dataset  $\mathcal{M}_1$ . The measurements are shown in Fig. 5. Note that  $\mathcal{M}_7$  contains more CAN messages than  $\mathcal{M}_1$ . However, a higher Hamming distance can be observed for  $\mathcal{M}_1$ . Specifically, 968 bits were flipped once or more while the vehicle was moving, whereas only 764 bits were flipped while the vehicle was stationary. Moreover, the sum of the Hamming distance differed;  $\approx 23\%$  of bit flips occurred more frequently while the vehicle was moving.

## V. EXPERIMENTAL RESULTS

The detection performance, feasibility, and explainability of X-CANIDS are discussed in this section. First, the parameters used in the experiment are described. Table II displays 62 streams and a maximum DLC of 8. Thus, we set  $N = 62$  and  $M = 64$  for the message receiver. Considering that the minimum and maximum values of the average time intervals were 0.01 s and 2 s, we initially assigned  $t = 0.01$ s and  $w = 200$  as baseline parameters. As noted in the previous section, there were 688 signal types. Thus, we initially achieved a  $200 \times 688$ -sized  $\mathbf{S}$  every 0.01 s.

We observed that X-CANIDS did not need to examine all signals in our vehicles to detect intrusions. Many static signals (e.g., the 10 signals of stream 5FAh in Table II) can be inspected using a simple rule, whether or not a change occurs. Moreover, certain signals contain checksums or sequential counters that are easily predictable. We excluded static signals and signals that contained the following keywords: *sum*, *alive*, *msgcount*, *msgcnt*, *paritybit*, and *mul\_code*. After excluding these signals, we obtained 107 signals from the 35 streams. Thus, the final feature shape of  $\mathbf{S}$  was  $200 \times 107$ .

### A. Parameters

1) *Autoencoder Layer*: First, we examine the optimal layer for  $\mathbf{AE}$ . We used six types of layers to implement the autoencoders and trained them for up to 2,000 epochs with an early-stopping patience of 50 epochs. The Adam optimizer fitted the models with a learning rate of 0.0001. A smaller MSE indicated better performance provided by a layer to  $\mathbf{AE}$ . The experimental results are presented in Fig. 6. The BiLSTM layer enabled the best performance in our experiment, followed by the LSTM and Conv2D layers.

2) *Feature Generation Parameters*: At the beginning of this section, we heuristically assigned  $t = 0.01$ s and  $w = 200$  as baseline parameters. Two parameters can affect the detection performance and process time of X-CANIDS. Therefore, two parameters need to be chosen carefully. While both are important factors, in this section, we explore optimum values

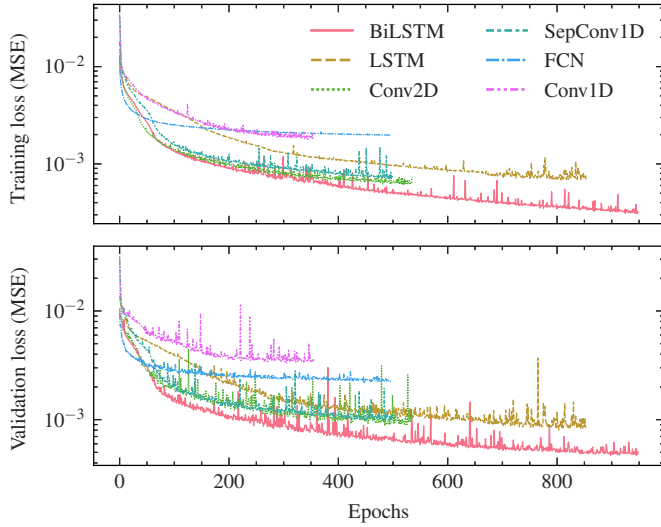


Fig. 6. Learning curves of six types of autoencoders. Each training run was terminated using the early-stopping strategy. The BiLSTM-based autoencoder exhibited the smallest reconstruction error of  $4.686 \times 10^{-4}$ , at epoch 948.

for these two parameters toward high intrusion detection performance. For the payload sampler (§III-B1), we consider the candidate time intervals as  $t \in \{0.001 \text{ s}, 0.002 \text{ s}, 0.005 \text{ s}, 0.01 \text{ s}, 0.02 \text{ s}, 0.05 \text{ s}, 0.1 \text{ s}\}$ . For the time-series feature generator (§III-B4), we consider the candidate window sizes as  $w \in \{25, 50, 75, 100, 150, 200, 300, 400\}$ . We prepared a test set with six attacks to compare the detection performances. The reader is referred to §V-F for further information regarding the attacks. As outputs from X-CANIDS are real numbers, we use the AUC—area under the receiver operating characteristic (ROC) curve, which was rendered while adjusting the detection threshold  $\Theta$ —as the primary evaluation metric.

Fig. 7 shows the validation losses and intrusion detection performances with the candidate parameters. For Fig. 7(a)–(d), we tested the candidate parameters with training sets from  $\mathcal{M}_3$ – $\mathcal{M}_4$ . The experimental results show a clear trend—the smaller the value we choose, the smaller the reconstruction error an AE provides. However, setting a very small number for two parameters might not be a good solution because a feature  $\mathbf{S}$  becomes representing a status of a very small time gap. Figures Fig. 7(b) and (d) support our concern that a too-small value shows a poor AUC score. Instead, in Fig. 7(b), we confirmed the best AUC of 0.975423 with  $t = 0.005 \text{ s}$ . Also, in Fig. 7(d), we confirmed the best AUC of 0.975440 with  $w = 150$ .

Fig. 7(e)–(f) compares two parameter combinations with the baseline parameters. In two figures, the entire training sets (i.e.,  $\mathcal{M}_1$ – $\mathcal{M}_4$ ) were used. Our baseline parameters ( $t = 0.01 \text{ s}$  and  $w = 200$ ) exhibited the moderate intrusion detection performance with the AUC of 0.9715. By changing  $t$  to 0.005 from 0.010, we confirmed the better AUC of 0.9838. We also confirmed the improved AUC of 0.9929 when we change  $w$  to 150 from 200 as well as  $t = 0.005 \text{ s}$ . Considering X-CANIDS is an unsupervised method, the performance that we confirmed would be acceptable. Therefore, we will utilize ( $t = 0.005 \text{ s}$  and  $w = 150$ ) for the rest of the paper.

TABLE III  
LAYOUT OF BiLSTM-BASED AE.

Layer	# parameters	Output shape	Symbol
Input	0	$150 \times 107$	$\mathbf{S}$
BiLSTM	184,040	$150 \times 214$	
BiLSTM	340,000	250	$h$
Repeat input 150 times	0	$150 \times 250$	
BiLSTM	306,448	$150 \times 214$	
BiLSTM	275,632	$150 \times 214$	
Time-distributed dense	23,005	$150 \times 107$	$\mathbf{S}'$

TABLE IV  
INTRUSION DETECTION PERFORMANCE AGAINST FUZZING ATTACKS.

Fuzzing rate	Bus load (%)	Precision	Recall	F1-score
10 msg./s	100.4584	0.999195	0.911456	0.953311
20 msg./s	100.9168	0.999251	0.992880	0.996055
30 msg./s	101.3752	0.999266	0.998621	0.998943
40 msg./s	101.8336	0.999261	0.999735	0.999498
50 msg./s	102.2920	0.999267	0.999927	0.999597
60 msg./s	102.7504	0.999262	0.999979	0.999620
70 msg./s	103.2088	0.999267	0.999969	0.999618
80 msg./s	103.6672	0.999194	0.999938	0.999566
90 msg./s	104.1256	0.999126	0.999974	0.999550
100 msg./s	104.5840	0.999121	0.999984	0.999553
200 msg./s	109.1681	0.999126	0.999984	0.999555
300 msg./s	113.7521	0.999121	0.999990	0.999555
400 msg./s	118.3362	0.998405	0.999990	0.999197
500 msg./s	122.9202	0.999023	0.999984	0.999503
1,000 msg./s	145.8404	0.998322	0.999995	0.999158
1,500 msg./s	168.7606	0.998322	0.999995	0.999158
2,000 msg./s	191.6808	0.998322	0.999995	0.999158

3) *Model*: Table III outlines the BiLSTM-based AE, which was used for the remainder of the experiments. The encoder compresses a  $150 \times 107$ -sized matrix  $\mathbf{S}$  into a 250-sized latent vector  $h$  (compression rate  $\approx 1.56\%$ ). The decoder reconstructs  $\mathbf{S}'$  using  $h$ . We assigned the parameter  $\Theta = 28.2$  (where  $q = 0.993$ ) for the intrusion detection.

### B. Intrusion Detection Performance

We tested the proposed method using the test set  $\mathcal{M}_6$ . We conducted attack simulations in the period 480–1440 s, half of the capture period of the test set, to obtain a label-balanced test set. We denoted the ground-truth labels as “attack” if a given  $\mathbf{S}$  was affected by the attack. As each input was generated every  $t = 0.005 \text{ s}$ , the detection result was also labeled every 0.005 s. We used precision, recall, and F1-score as the performance evaluation metrics. Our evaluation strategy was to simulate multiple attacks on the designated period and feed each dataset to X-CANIDS. Regarding the fuzzing attack, we built 17 intrusion datasets with various fuzzing rates. Regarding the fabrication, masquerade, and suspension attacks, we tried each attack on every single stream that contributed to the feature creation. To this end, we built 105 intrusion datasets. Regarding the replay attack, we built four intrusion datasets with different capture durations. Consequently, we conducted 126 experiments with different datasets.

1) *Fuzzing*: Our evaluation starts with the fuzzing attack. As discussed in §II-D, the adversary injects CAN messages with random  $a$  and  $p$ . The adversary defines  $a$  with one of AIDs listed in Table II. We also tried various fuzzing rates. A fuzzing rate means the number of injected CAN messages

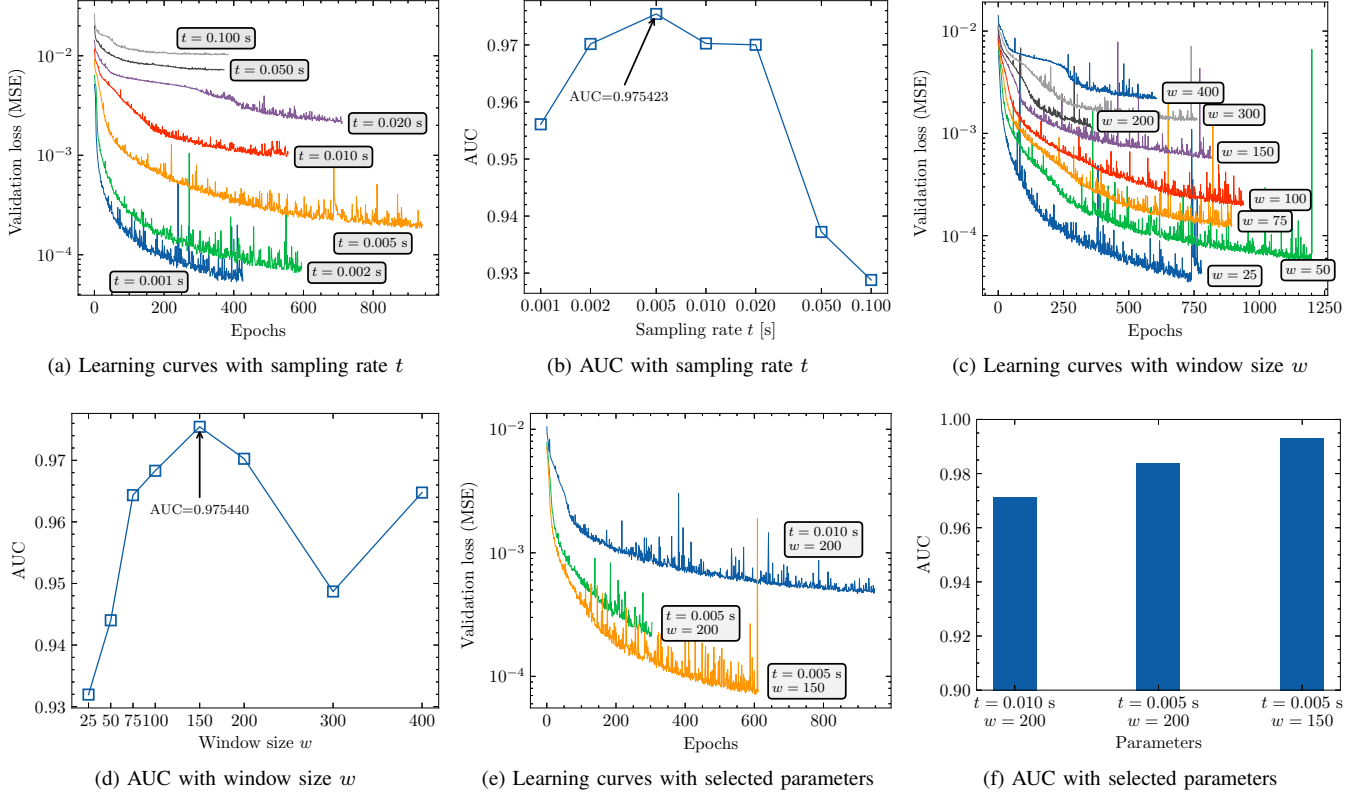


Fig. 7. Validation losses and AUCs with parameters. In (a), (c), and (e), smaller values result in smaller validation losses. On the other hand, two parameters ( $t = 0.005$  s,  $w = 150$ ) showed the best intrusion detection performances in (b) and (d), respectively. Using two values exhibited the AUC of 0.9929 in (f).

per s. Table IV presents the intrusion detection performance against fuzzing attacks with various fuzzing rates. The bus load column indicates the change in the number of CAN messages during the fuzzing attack. Note that the average number of CAN messages per second was  $\approx 2181.48$  in the CAN bus. It can be observed that the overall detection performance was outstanding. In particular, X-CANIDS distinguished small-scale attacks with a fuzzing rate of 10 messages per second. Even when 0.4584% of the bus load increased compared with the attack-free state, the proposed method achieved a recall of 0.911456. As the fuzzing rate increased, X-CANIDS identified nearly all intrusions, with a recall of  $\approx 1$ . The precision was always higher than 0.998, indicating a false positive rate of less than 0.002.

2) *Fabrication, Masquerade, and Suspension*: Table V presents the intrusion detection performance against the fabrication, masquerade, and suspension attacks. Regarding the fabrication and masquerade attacks, X-CANIDS showed outstanding performances with the F1-score  $\geq 0.99$ , except for four streams—044h, 381h, 5A0h, and 5B0h. Even though X-CANIDS successfully identified intrusions on streams 044h, 5A0h, and 5B0h with high recalls  $\geq 0.992$ , smaller precision confirmed that there were some false positive cases. Meanwhile, X-CANIDS did not detect intrusions on stream 381h. Based on average scores, we can rely on X-CANIDS to protect a vehicle from potential fabrication and masquerade attacks.

Our experimental results confirmed that X-CANIDS is less effective for suspension attacks. Although we could achieved

the high precision scores, small recall scores imply that X-CANIDS missed suspension attacks. The proposed method exhibited poor performance because a suspension of a certain stream does not compromise values in the buffer  $\mathbf{P}$ . As a result,  $\mathbf{P}_{\text{copy}}$  will contain legitimate values even during a suspension attack. Fortunately, considering the nature of the suspension attack, we can detect a suspension attack easily through a measurement of the number of messages in each stream. Since the precision scores are moderate, We can utilize X-CANIDS as a secondary network monitor.

3) *Replay*: We conducted four further experiments to evaluate the detection performance against replay attacks. Table VI lists the capture durations for these experiments. For each experiment, an adversary captures a series of all legitimate CAN messages broadcasted on a CAN bus. Then, the adversary replays the series repeatedly during the designated attack period (*i.e.*, 480–1440 s). For instance, in the first experiment, the content of the replayed series corresponds to a subset of the test set  $\mathcal{M}_6$  in the period of 0–120 s. All streams were captured in the series. During the replay attack, the number of transmitted CAN messages per second doubled. However, the injected payloads originated from legitimate ECUs. The results in Table VI demonstrate that X-CANIDS is effective against replay attacks even though legitimate ECUs in our vehicle have generated all injected payloads.

TABLE V  
INTRUSION DETECTION PERFORMANCE AGAINST FABRICATION, MASQUERADE, AND SUSPENSION ATTACKS.

Target AID	Fabrication attack			Masquerade attack			Suspension attack		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
044h	0.755961	0.999993	0.861020	0.755971	0.999993	0.861027	0.707935	0.038682	0.073356
080h	0.999267	0.999979	0.999623	0.999262	1.000000	0.999631	0.998766	0.594153	0.745073
081h	0.999262	1.000000	0.999631	0.999267	0.999984	0.999625	0.999085	0.801305	0.889332
111h	0.999267	0.999979	0.999623	0.999267	0.999995	0.999631	0.999034	0.758901	0.862566
112h	0.999267	0.999979	0.999623	0.999267	1.000000	0.999633	0.999111	0.824626	0.903521
113h	0.999267	0.999979	0.999623	0.999267	0.999995	0.999631	0.998593	0.520752	0.684531
162h	0.999267	0.999979	0.999623	0.999267	0.999995	0.999631	0.998267	0.422766	0.593981
18Fh	0.999267	0.999979	0.999623	0.999262	1.000000	0.999631	0.998264	0.421860	0.593086
200h	0.999262	1.000000	0.999631	0.999267	0.999984	0.999625	0.982244	0.040593	0.077965
220h	0.999267	0.999979	0.999623	0.999262	1.000000	0.999631	0.982360	0.040865	0.078465
251h	0.999267	0.999922	0.999594	0.999267	0.999922	0.999594	0.982237	0.040578	0.077937
260h	0.999262	1.000000	0.999631	0.999267	0.999984	0.999625	0.998275	0.424670	0.595860
2B0h	0.999267	0.999818	0.999542	0.999267	0.999886	0.999576	0.982228	0.040557	0.077898
316h	0.999267	0.999984	0.999625	0.999256	1.000000	0.999628	0.999098	0.818234	0.899666
329h	0.999267	1.000000	0.999633	0.999262	1.000000	0.999631	0.999147	0.859115	0.923854
381h	0.982246	0.040600	0.077976	0.982246	0.040600	0.077976	0.982246	0.040600	0.077976
383h	0.999256	0.999995	0.999625	0.999251	1.000000	0.999625	0.982235	0.040574	0.077928
386h	0.999256	1.000000	0.999628	0.999251	0.999979	0.999615	0.999044	0.778066	0.874816
387h	0.999266	0.999630	0.999448	0.999266	0.999662	0.999464	0.980727	0.037341	0.071943
47Fh	0.999266	0.999365	0.999316	0.999266	0.999521	0.999394	0.982657	0.041578	0.079781
4F1h	0.999262	0.999979	0.999620	0.999251	1.000000	0.999625	0.999009	0.750446	0.857070
50Ch	0.999178	0.999995	0.999586	0.999267	0.999901	0.999584	0.982237	0.040584	0.077947
52Ah	0.999074	0.999792	0.999433	0.999064	1.000000	0.999532	0.998850	0.809012	0.893964
541h	0.999267	0.999938	0.999602	0.999267	0.999740	0.999503	0.998442	0.470205	0.639326
545h	0.999173	0.999995	0.999584	0.999173	0.999995	0.999584	0.981965	0.039957	0.076789
547h	0.999178	0.999984	0.999581	0.999199	0.999979	0.999589	0.996630	0.217033	0.356444
549h	0.999173	0.999995	0.999584	0.999168	1.000000	0.999584	0.982251	0.040612	0.078000
553h	0.999266	0.999172	0.999219	0.999266	0.998751	0.999008	0.998449	0.472302	0.641263
555h	0.999168	0.999995	0.999581	0.999173	0.999990	0.999581	0.982217	0.040536	0.077859
556h	0.999168	1.000000	0.999584	0.999173	0.999984	0.999579	0.999010	0.835478	0.909955
557h	0.999209	0.999932	0.999571	0.999173	0.999995	0.999584	0.986437	0.053375	0.101270
58Bh	0.999267	0.999984	0.999625	0.999178	0.999891	0.999534	0.978268	0.033034	0.063911
593h	0.999266	0.999755	0.999511	0.999085	0.999984	0.999534	0.978354	0.041413	0.079462
5A0h	0.822140	0.992078	0.899150	0.822323	0.994221	0.900139	0.748007	0.041658	0.078921
5B0h	0.749149	0.999951	0.856569	0.749131	1.000000	0.856575	0.727525	0.040083	0.075980
Average	0.979597	0.972277	0.962327	0.979596	0.972341	0.962353	0.968263	0.328901	0.407648

TABLE VI  
INTRUSION DETECTION PERFORMANCE AGAINST REPLAY ATTACKS.

Capture duration (s)	Precision	Recall	F1-score
0–120	0.999266	0.998954	0.999110
120–240	0.999178	0.999990	0.999584
240–360	0.999204	0.999781	0.999493
360–480	0.999262	0.993266	0.996255

### C. Performance Comparison with Prior Research

As it is difficult for researchers to obtain CAN databases, previous payload-based studies [29], [30], [13], [12], [31] used the raw payloads of CAN messages as inputs. For a comparison of the detection performance with prior research, we implemented a method known as CANnolo proposed by Longari *et al.* [31] because the concept of the study is similar to that of our method. They proposed a self-supervised IDS using an LSTM-based autoencoder. CANnolo is supposed to be trained with benign time-series payloads in the bit representation. Consequently, we trained the model using our training set.

Fig. 8 depicts five receiver operating characteristic (ROC) curves, each of which was rendered while adjusting the detection threshold  $\Theta$ . Curves 1 and 2 represent the detection performances of X-CANIDS and CANnolo, respectively. X-

- 1. Proposed method (AUC = 0.9929)
- - - 2. CANnolo (Longari *et al.*) (AUC = 0.8993)
- ⋯ 3. CANnolo + BiLSTM (AUC = 0.9517)
- · - 4. Trained w/ stationary set  $\mathcal{M}_7$  (AUC = 0.6100)
- - - 5. Trained w/ stationary set  $\mathcal{M}_7$ , w/o CAN DB (AUC = 0.4699)

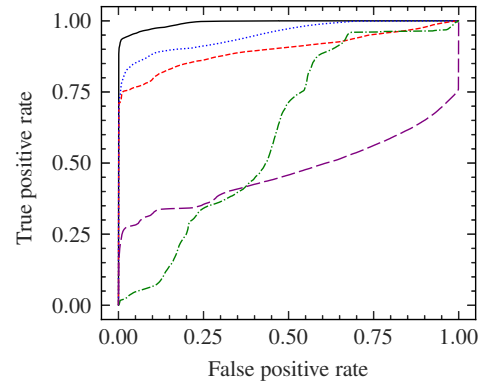


Fig. 8. ROC curves. X-CANIDS (curve 1) exhibited better intrusion detection performance compared to the previous work (curve 2) [31].

CANIDS was superior to CANnolo, with a 0.0936 gain in the area under the curve (AUC). Furthermore, we revised

CANnolo by replacing the LSTM layers with BiLSTM layers as BiLSTM exhibits a smaller reconstruction error than LSTM (see Fig. 6). A comparison of curves 2 and 3 reveals that the BiLSTM layer helped CANnolo to detect anomalies more accurately. However, X-CANIDS still performed better than the revised CANnolo. We conclude that the signals that were deserialized from the raw payloads were more helpful in detecting anomalies.

#### D. Advantages of Using Driving Dataset and Signals

The proposed method should be trained with a driving dataset rather than a stationary or simulated dataset to ensure high detection performance. We conducted further experiments with a stationary dataset to confirm the importance of using the driving dataset in the training phase. In Fig. 8, curve 4 indicates the performance of X-CANIDS when it was trained with  $\mathcal{M}_7$ . A comparison of curves 1 and 4 demonstrates that the use of the driving dataset significantly improved the detection performance.

Moreover, we assume that it is necessary to train **AE** with raw payloads, owing to the lack of a CAN database. Curve 5 shows the detection performance under this assumption. A comparison of curves 4 and 5 confirms that the signals aided in achieving better detection performance.

#### E. Feasibility Consideration

It is necessary to ensure that no bottleneck occurs owing to the computation time. The computation time is dependent on the complexity of the method and the computational power that is provided by an in-vehicle component. We selected an NVIDIA Jetson AGX Xavier, which is an automotive-grade embedded device equipped with a GPU and CAN shield, to investigate the feasibility. This device is plausible because it is used to compute automotive applications in both the vehicle industry and academia (e.g., [32]). We implemented X-CANIDS on the device. The device could monitor the CAN bus in the vehicle using the CAN shield. The GPU allowed us to compute  $f(\cdot)$  concurrently with a small batch of features. Owing to the device supporting TensorFlow natively, we could port the pretrained weights of  $f(\cdot)$  from our PC.

1) *Throughput*: We first measure the throughput of our **AE** on the device. Note that we chose  $t = 0.005$  s. It means that the feature generator generates 200 features per s. Consequently, the throughput must be equal to or greater than 200 samples per s. We tried 11 batch sizes and summarized the result in Table VII. We denote the batch size as  $B$ . We confirmed that the device can process up to  $\approx 2,010$  samples per s when we use  $B = 1024$ . In order to minimize the batch completion time as well as the detection latency, we decide the  $B = 8$ .

2) *Detection latency*: The detection latency is the time gap between the attack and alert. When  $B = 8$  was selected, the

TABLE VII  
THROUGHPUT AND INFERENCE TIME ON NVIDIA JETSON AGX XAVIER.

Batch size $B$	Throughput (samples/s)	Inference time $t_\beta$ (ms/sample)
4	114.6033	8.7258
8	239.1564	4.1814
16	476.0859	2.1005
32	870.1233	1.1493
64	1,257.3287	0.7953
128	1,526.7854	0.6550
256	1,773.3477	0.5639
512	1,928.1721	0.5186
1,024	2,010.5995	0.4974
2,048	1,881.2511	0.5316
4,096	1,234.2686	0.8102

detection latency was derived as follows:

$$\begin{aligned}
 \text{Detection latency} &= zt + t_\alpha + Bt_\beta \\
 &= z \cdot 5 \text{ ms} + 4.8 \text{ ms} + 8 \cdot 4.1814 \text{ ms} \\
 &= \begin{cases} 38.2512 \text{ ms,} & \text{if } z = 0, \text{ the batch is full on arrival.} \\ 73.2512 \text{ ms,} & \text{if } z = 7, \text{ the batch is empty.} \end{cases} \quad (9)
 \end{aligned}$$

In the above,  $t_\alpha = 4.8$  ms stands for the time consumption that the feature generator takes, and  $t_\beta$  stands for the inference time per sample (see Table VII). Also,  $z \in \{0..(B-1)\}$  is the number of inputs required to complete a batch; thus,  $zt$  represents the batch completion time. We conclude that X-CANIDS provides an intrusion alert with a deterministic latency of no greater than 73.2512 ms.

The deterministic latency is high compared to the minimum time intervals of existing streams in our vehicle (i.e., 10 ms, see Table II). Carmakers considering installing an intrusion response system (IRS) should note that while X-CANIDS can be used for intrusion detection, it cannot be seamlessly integrated with their IRS to mitigate identified cyberattacks in time without further optimizing the detection latency. We have deferred the optimization to future work.

3) *CPU, RAM and GPU usage*: We utilized the embedded device only to evaluate the feasibility of X-CANIDS. However, the device may run other processes along with X-CANIDS in a real-world scenario. For those who are considering the scenario, we measured the CPU and RAM usage on our device. Our inference program—implemented using C++ with standard template library—takes the CPU utilization of  $\approx 150\%$  (100% per core, 8 cores total) and the memory space of  $\approx 2.7$ GB including deep learning libraries. We expect the CPU and RAM requirements steady in other environments because X-CANIDS does not rely on sparse data (e.g., one-hot vector) or dynamic memory allocation. The bus load of a CAN-based IVN does not affect the performance either because the feature generator builds an **S** every designated  $t$  s. We had no choice but to use two threads since  $t_\alpha = 4.8$  ms was nearly equal to the feature creation interval  $t$ ; one thread generates features, and another thread deals with the rest—the batch compilation, feeding a batch to the model, and measure the error rate. Finally, the GPU utilization was measured as  $\approx 63\%$  on average when  $B = 8$ .

### F. Explanation of Detection Results

In this section, we discuss the explainability of our proposed framework. A heatmap that plots the error rates of the 107 signals over time is presented in Fig. 9. We conducted six attacks on our test set and measured their error rates. The figure shows only the error rates that exceeded the detection threshold  $\Theta = 28.2$ . That is, the marked points in the figure represent the predicted intrusions. The attack period and description are indicated at the top of the heatmap. The moving average of the detection accuracy is also provided.

1) *Period 1—Fuzzing*: An intensive fuzzing attack was conducted during 480–720 s and the error rate exceeded the detection threshold for almost all signals. The  $\text{argmax}(\mathbf{r})$  function continuously pointed out signal 5B0\_CF\_Clu\_Odometer. Nevertheless, an expert who is responsible for incident response would be able to identify the type of attack as fuzzing because of the multiple simultaneous errors. Unfortunately, during this period, the proposed framework did not work for the following signals: 112\_VS\_TCU\_DECIMAL, 220\_YAW\_RATE, and 4F1\_CF\_Clu\_VanzDecimal.

2) *Period 2—Fabrication*: We considered an adversary who attempts to fabricate a portion of the payloads for signal 556\_PID\_0Ch during 840–960 s. The signal is part of the OBD-II freeze frame containing the parameter ID 0Ch. That is, the signal represents the current engine RPM [33]. The experimental results showed that 556\_PID\_0Ch reached the highest error rate of  $\approx 10^4$  during this period. Three signals representing the RPM, namely 080\_N, 162\_Cluster\_Driving\_RPM, and 316\_N, also exhibited high error rates.

3) *Period 3—Fabrication*: An adversary who injects CAN messages with  $a=162h$  was assumed during 1080–1200 s. We can see that three signals belonging to stream 162h exhibited high error rates. In particular,  $\text{argmax}(\mathbf{r})$  successfully pointed out signal 162\_Cluster\_Enging\_RPM.

4) *Period 4—Masquerade*: We considered an adversary who attempts to report a fraudulent current velocity to the driver via an instrumental cluster. For this purpose, we simulated a masquerade attack that changes signal 316\_VS during 1320–1440 s. As shown in the figure, the signal exhibited the highest error rate. Furthermore, other speed-related signals exhibited high error rates simultaneously, including the four wheel speed signals that are defined in stream 386h.

5) *Period 5—Suspension*: We performed a suspension attack for stream 556h during 1560–1680 s. As demonstrated in Table V, the proposed method is ineffective against suspension attacks. The figure also supports the weakness of X-CANIDS. We observed that the detection accuracy increased and decreased during this period. The proposed framework identified anomalies in signal 556\_PID\_0Dh, which represents the current velocity (see the description of parameter ID 0Dh in [33]). The error rate of the signal exceeded the detection threshold when the vehicle velocity differed from 68 km/h (cf., the red-filled area in the velocity chart and change in the moving accuracy).

6) *Period 6—Fabrication*: Finally, motivated by previous works [15], [24], we considered a max coolant temperature attack. In our dataset, signal 329\_TEMP\_ENG deals with the engine coolant temperature. Thus, we conducted a fabrication

TABLE VIII  
PREVIOUS WORK THAT PROPOSED IN-VEHICLE IDSS FOR CAN BUSES.

Research	Timestamp	AID sequence	Payload	Signal	Output	Inference time
Müter and Asaj [34]	✓		✓		Binary	—
Kang and Kang [29]			✓		Binary	2–5 ms <sup>†</sup>
Marchetti and Stabili [35]		✓			Binary	—
Taylor <i>et al.</i> [36]	✓		✓		Binary	—
Song <i>et al.</i> [25]	✓				Binary	1 ms <sup>†</sup>
Taylor <i>et al.</i> [30]			✓		Binary	—
Marchetti <i>et al.</i> [37]	✓				Binary	—
Stabili <i>et al.</i> [38]			✓		Binary	—
Markovitz and Wool [22]				✓	Binary	—
Wasicek <i>et al.</i> [14]				✓	Real number	—
Tomlinson <i>et al.</i> [26]	✓				Binary	—
Olufowobi <i>et al.</i> [27]	✓				Binary	9–10 ms <sup>†</sup>
Young <i>et al.</i> [28]	✓				Binary	—
Katragadda <i>et al.</i> [39]		✓			Binary	151 ms <sup>‡</sup>
Song <i>et al.</i> [40]		✓			Binary	5–6.7 ms <sup>†</sup>
Longari <i>et al.</i> [31]			✓		Binary	—
Hossain <i>et al.</i> [12]		✓	✓		Category	—
Tariq <i>et al.</i> [13]	✓	✓	✓		Category	14–73 ms <sup>†</sup>
Song and Kim [41]		✓			Binary	—
Shahriar <i>et al.</i> [15]				✓	Real number	—
Hoang and Kim [42]		✓			Binary	0.63 ms <sup>†</sup>
X-CANIDS (this work)				✓	Real number	4.18 ms <sup>‡</sup>

<sup>†</sup>per CAN message

<sup>‡</sup>per feature

\*measured on an embedded device

attack on the signal. It can be observed that  $\text{argmax}(\mathbf{r})$  successfully pointed out the exact target signal during this period.

## VI. RELATED WORKS

In this section, related works on in-vehicle IDSSs for CAN buses are reviewed. Table VIII lists 21 previous studies, with the input type used to detect the intrusion, output type, and inference time presented in the respective paper. In many cases, an IDS returns a binary for each input to indicate whether the vehicle is attacked. Meanwhile, two studies [12], [13] proposed IDSSs that return a categorical value. This category refers to the type of attack. These IDSSs require labeled training sets to recognize the attack type. Two signal-aware IDSSs ([14], [15]) return a real number that can be used as the detection result and confidence score.

The primary goal of an in-vehicle IDS is to detect anomalies effectively. Moreover, an IDS should be implemented on an ECU or embedded device that provides limited computational power. Therefore, IDSSs should be tested on these devices to confirm their feasibility. In particular, the inference time should be compared with the feature-creation frequency. However, excluding this work, only seven studies measured the inference time. Moreover, none of the studies mentioned that the inference time was measured using an embedded device.

An in-vehicle IDS assesses the in-vehicle traffic using the timestamp, AID sequence, payload, and/or signal. In the remainder of this section, we categorize the related studies according to the input type.

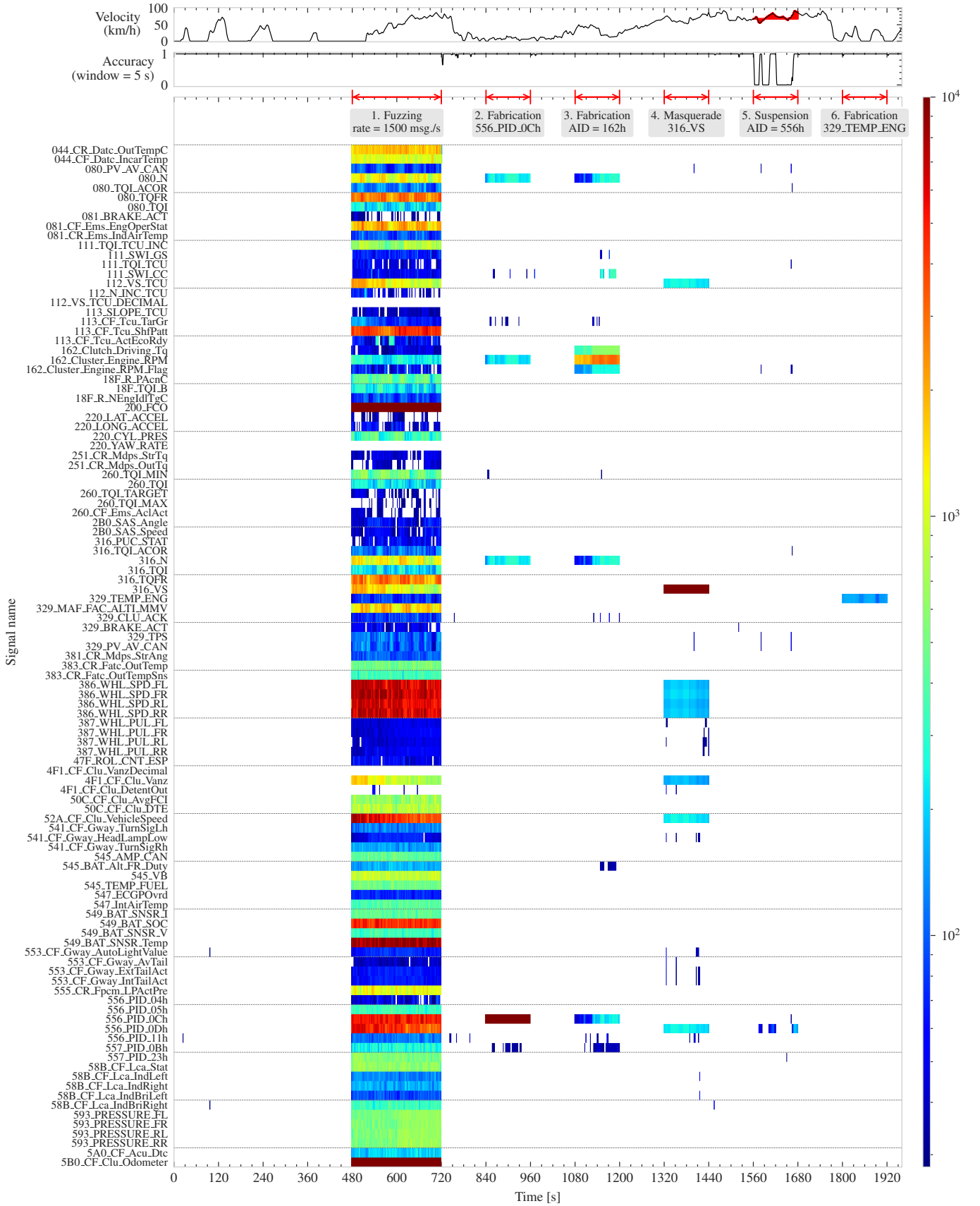


Fig. 9. Error rates over test dataset  $\mathcal{M}_6$ . The error rates are depicted as  $28.2 = \Theta \leq \text{blue} \leq 10^2 \leq \text{green} \leq 10^3 \leq \text{red} \leq 10^4 +$ . We conducted attacks in six periods. In attack period 1, the error rates were exceeded on many signals owing to intensive fuzzing. In attack periods 2–6, we could distinguish the exact target signals showing a maximum error rate at a certain time. The reader is referred to the color version of this page for interpretation of the figure.

### A. Time Interval-Based IDS

A timestamp is not an officially supported field in the CAN frame. Nevertheless, an ECU can measure the relative time of the transmission itself because every ECU that shares the same medium is time synchronized bit by bit. ECUs tend to report their status periodically. Time-interval-based IDSs exploit these mechanisms. The measurement of timestamps is an efficient means of detecting anomalies. This also enables the detected intrusions to be understood. In contrast, a time-interval-based IDS does not work for the sporadic transmission of CAN messages.

Most early studies leveraged the periodic transmission mechanism. For example, in 2011, Müter and Asaj [34] explored the applicability of the entropy-based anomaly detection method. The proposed method takes advantage of the periodicity of IVN traffic. They measured the normal probability distribution of attack-free CAN data and compared it with that of abnormal CAN data. However, the study lacked an explicit threshold determination method or evaluation result, such as the detection accuracy. In 2016, Song *et al.* [25] proposed a rule-based intrusion detection method that measures the time interval of two adjacent CAN messages. When the time interval of a new CAN message is shorter than the threshold, the proposed IDS considers the message as an intrusion. The experimental results demonstrated high detection performance with handcrafted thresholds. However, the method for determining the threshold remains to be improved. Marchetti *et al.* [37] proposed an entropy-based anomaly detector for CAN buses. Tomlinson *et al.* [26] adopted the ARIMA model and Z-score to identify CAN message timing anomalies in a time window. Olufowobi *et al.* [27] proposed an intrusion detection method known as SAIDuCANT, which leverages the periodic message behavior in the CAN bus. Young *et al.* [28] adopted fast Fourier transform to measure CAN message update frequency in a stream.

### B. Sequence-Based IDSs

Sequence prediction is a well-known machine-learning problem. A sequence predictor predicts the next symbol based on previously observed categorical data. A sequence of AIDs can be used as sequence predictors to model the attack-free state of IVNs for in-vehicle traffic monitoring. Sequence-based IDSs can be used in all types of CAN-based IVNs because it is easy to compile AID sequences. However, these methods exhibit several drawbacks. Particularly, it is difficult to understand why a given sequence is classified as abnormal. Moreover, an adversary may falsify a legitimate sequence by conducting an intensive replay attack.

Marchetti and Stabili [35] compiled a transition matrix to represent the recurring patterns of two adjacent AIDs. The transition matrix is used to evaluate the CAN bus stream in the inference phase. Katragadda *et al.* [39] proposed a message sequence-based anomaly detection method that builds a frequent sequence tree, where each node represents a subsequence and each edge measures an observation count. The sequence length should be carefully selected because detection performance and feasibility are dependent thereon.

Tariq *et al.* [13] combined the benefits of rule-based models and neural networks. They claimed that their heuristic model works for known attack signatures, whereas neural networks can cope with unknown attacks. This is the only work that simultaneously examined the timestamp, payload, and AID sequences. Indeed, their method requires a substantially longer inference time than those of other approaches. Song *et al.* [40] proposed a CNN that is a variation of Inception-ResNet to examine AID sequences. Their detection model outperformed conventional detection methods. However, a limitation of the proposed method is that it requires a labeled dataset to train the model. To overcome this drawback, they proposed another training method [41] to train their CNN model in an unsupervised manner. Hoang and Kim [42] proposed an adversarial autoencoder that attempts to reconstruct  $29 \times 29$ -sized features, which represent 29 continuous AIDs in bits.

### C. Payload-Based IDSs

Payload-based IDSs evaluate the payloads in CAN messages, where each payload is represented as bit sequences. The training set needs to be carefully prepared to use payload-based IDSs in real-world scenarios; otherwise, the payload dynamics could cause false alarms.

Taylor *et al.* [36] measured the number of packets and average Hamming distance of the CAN message payloads in a sliding window. The statistical features that were derived from these two values were used to train a one-class support vector machine. Their experimental results were dependent on the window size for feature generation. Unfortunately, the unsupervised method exhibited a considerable false positive rate with the incorrect window size. Kang and Kang [29] proposed a binary intrusion detection method based on a fully connected neural network, which uses 64-dimensional bit sequences (*i.e.*, a payload of CAN messages) and then returns a logistic value of 0 or 1. They used a packet generator known as OCTANE to evaluate the proposed method. Despite the high performance of the experimental results, this work exhibits two limitations: (1) the method was evaluated using only three streams and (2) the simulated payloads that were used in the experiment may not reflect real-world situations. Taylor *et al.* [30] proposed an intrusion detector that consists of LSTM-based models for each stream. Each model uses a  $20 \times 64$ -sized matrix that comprises 20 subsequences of CAN payloads in the bit representation. Subsequently, the model predicts the next payload of a stream. They considered five loss metrics to measure the anomaly score. Stabili *et al.* [38] measured the Hamming distance for each stream. They classified each stream into the no distance, small distance, and mid-distance ranges. When a stream exceeds a given distance range in the inference phase, it is classified as an anomaly. However, the method was evaluated using only fuzzing attacks. Hossain *et al.* [12] developed a supervised IDS using an LSTM model that considers the raw payloads of a CAN message. Longari *et al.* [31] deployed an anomaly detection system based on LSTM autoencoders. They implemented an LSTM model for each stream to reconstruct a time-series bit sequence of payloads.

#### D. Signal-Aware IDSs

To the best of our knowledge, Markovitz and Wool [22] were the first to use portions of the payload in CAN messages to detect intrusions, as opposed to using an entire payload. They divided a 64-bit payload into several fields, each of which was assigned as a constant, categorical, counter, or sensor type. A rule-based detection algorithm was considered for each data type. Wasicek *et al.* [14] proposed an IDS that uses 54 types of signals that are obtained via OBD-II PIDs. They reported that a fully connected network with a bottleneck can be used to measure the anomaly score (*i.e.*, the reconstruction error). In 2022, Shahriar *et al.* [15] proposed a signal-aware IDS, named CANShield. A 2D CNN was designed to reconstruct the signals that were deserialized from raw payloads. They used CAN-D [16], which is an automatic dissector for CAN traffic, to obtain these signals.

#### E. Comparison with related works

In the literature, two related works [14], [15] and X-CANIDS are closely aligned in that “an autoencoder takes signals to detect in-vehicle intrusion.” In this section, we provide a comparative analysis to discuss the advantages of X-CANIDS in terms of signal processing and the detection model.

**Signal process.** The method in [14] takes signals from OBD-II responses, while X-CANIDS and CANShield [15] use signals from CAN messages. OBD-II responses are an alternative source to obtain signals without a CAN database. However, these signals primarily reflect powertrain-related sensor values, as the main objective of OBD-II is vehicle diagnosis. As a result, the method in [14] is insufficient to protect non-powertrain applications. Another limitation is that an OBD-II response will yield compromised signals once the powertrain has been compromised. Therefore, signals from CAN messages are more beneficial as they can reflect attempted attacks instantly. Meanwhile, CANShield has two limitations in signal processing. First, the method only analyzes a few pre-selected signals while ignoring the rest by design. Second, the order of signals can affect learning efficacy since the signals are converted into a 2D image. Compared to these related works, X-CANIDS has the following advantages: (1) it uses signals from CAN messages, (2) it can analyze all available signals, and (3) it ensures robust learning efficacy regardless of the order of signals.

**Model.** The methods in [14], [15] use an FCN- and a Conv2D-based autoencoder, respectively. In addition to these types of layers, we have tested six different layers to implement autoencoders. Our experimental results, using real driving datasets, show that the BiLSTM layer is superior to the FCN and Conv2D layers for the signal reconstruction problem (refer to §V-A1 and Fig. 6). Investigating the most effective layer is also a key contribution of this work.

### VII. DISCUSSION AND CONCLUSIONS

#### A. Limitation

We here discuss three limitations as well as remediation strategies for them. First, we used only one vehicle to evaluate

X-CANIDS. We found it difficult to arrange another car owing to the lack of a CAN database. To tackle the issue, we encourage carmakers to share their CAN databases with academia for research purposes. Otherwise, providing an application programming interface that allows researchers to obtain signals can be an alternative solution.

Second, X-CANIDS is insufficient for detecting suspension attacks. To this end, our future work will expand the proposed method to examine time-interval-based features along with signal features. Except for the masquerade attack, the rest of the attacks can be conducted by injecting or suppressing CAN messages. Therefore, analyzing a transmission period per stream would be a promising approach. The transmission period can also be considered an intuitive explanation because an expert can easily compare a current one with an expected one.

Finally, X-CANIDS takes some time to process the input and raise the alarm. Many streams have an update interval of 10 ms. However, we confirmed the detection latency of 73.2512 ms in the worst case. It implies there is a time gap exposed to the attack without awareness. To tackle the issue, we could get prompt detection results by considering a lightweight autoencoder or dedicated hardware that accelerates the autoencoder. IDSs with substantial computational demands may not be appropriate for all vehicle tiers, from low- to high-end. To achieve real-time detection at a reasonable cost, there is a compelling need for more lightweight IDS mechanisms.

#### B. Remarks and Conclusion

Recent vehicles are driven by software and have a broad attack surface. So far, many studies have been proposed for the precise detection of intrusions on in-vehicle networks. However, due to the lack of information on payload serialization, only a few studies have considered analyzing signals of CAN messages. Also, feasibility considerations are lacking. In response, cybersecurity regulations, including UNR 155, enforce the installation of an IDS inside vehicles and the analysis of cyberattacks. Therefore, the feasibility and explainability of in-vehicle IDSs are important for vehicle industries.

In this study, we have proposed X-CANIDS, in which the feature generator is designed to process live streams and create a time-series representation of the signals. A CAN database is combined with X-CANIDS to deserialize the signals from the CAN message payloads. We tested six types of autoencoders. The LSTM layer has often been considered in the literature [30], [13], [12], [31]. A Conv2D-based autoencoder has also been employed to model signals [15]. However, we demonstrated that the BiLSTM-based autoencoder outperformed the LSTM and 2D-CNN autoencoders. Then, we explored two parameters regarding feature generation. X-CANIDS expects an onboard AI-inference device to leverage the model. In case of a lack of such a device, a driver may consider installing a device like an after market autonomous driving kit of Comma.ai.

In summary, the experimental results suggest that X-CANIDS detects zero-day intrusions that are not observed during the training phase. In particular, the proposed method

offers an advantage for masquerade attacks that cannot be detected by time-interval- or sequence-based IDSs. X-CANIDS also offers outstanding performance against fuzzing, fabrication, and replay attacks.

We have considered the feasibility and explainability. To the best of our knowledge, these characteristics have not been considered in the previous works for CAN IDSs. Our feasibility evaluation confirms that X-CANIDS is able to be implemented on an embedded device to monitor live in-vehicle traffic while driving. The explainability provides an explicit hint about target ECUs or compromised signals. The explainability of X-CANIDS will help incident response teams analyze conducted cyberattacks. For that, X-CANIDS requires a CAN database. We claim that our method will be valuable to all carmakers because they can access CAN databases for their vehicles.

## REFERENCES

- [1] K.-T. Cho and K. G. Shin, "Error handling of in-vehicle networks makes them vulnerable," in *Proc. ACM CCS '16*, 2016, pp. 1044–1055.
- [2] H. Lee, S. H. Jeong, and H. K. Kim, "OTIDS: A novel intrusion detection system for in-vehicle network by using remote frame," in *Proc. PST '17*, 2017.
- [3] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," in *Proc. Black Hat USA '15*, Aug. 2015, pp. 1–91.
- [4] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Computers & Security*, vol. 103, p. 102150, 2021.
- [5] H. J. Jo, W. Choi, S. Y. Na, S. Woo, and D. H. Lee, "Vulnerabilities of Android OS-based telematics system," *Wireless Personal Communications*, vol. 92, no. 4, pp. 1511–1530, 2017.
- [6] S. Gayou, "Jailbreaking Subaru StarLink," <https://github.com/sgayou/subaru-starlink-research>, Nov. 2018.
- [7] "Tesla car hacked at Pwn2Own contest," <https://www.zdnet.com/article/tesla-car-hacked-at-pwn2own-contest/>, Mar. 2019.
- [8] G. Costantino and I. Matteucci, "CANDY CREAM - hacking infotainment Android systems to command instrument cluster via CAN data frame," in *Proc. IEEE CSE '19*, 2019, pp. 476–481.
- [9] Tencent Keen Security Lab, "Mercedes-Benz MBUX security research report," [https://keenlab.tencent.com/en/whitepapers/Mercedes-Benz\\_Security\\_Report\\_Final.pdf](https://keenlab.tencent.com/en/whitepapers/Mercedes-Benz_Security_Report_Final.pdf), Tech. Rep., May 2021.
- [10] H. J. Jo and W. Choi, "A survey of attacks on Controller Area Networks and corresponding countermeasures," *IEEE Transactions on Intelligent Transportation Systems*, pp. 6123–6141, Jul. 2022.
- [11] "UN Regulation No. 155 - cyber security and cyber security management system," E/ECE/TRANS/505/Rev.3/Add.154, Apr. 2021.
- [12] M. D. Hossain, H. Inoue, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based intrusion detection system for in-vehicle CAN bus communications," *IEEE Access*, vol. 8, pp. 185 489–185 502, 2020.
- [13] S. Tariq, S. Lee, H. K. Kim, and S. S. Woo, "CAN-ADF: The Controller Area Network attack detection framework," *Computers & Security*, vol. 94, p. 101857, 2020.
- [14] A. Wasicek, M. D. Pesé, A. Weimerskirch, Y. Burakova, and K. Singh, "Context-aware intrusion detection in automotive control systems," in *Proc. 5th ESCAR USA '17*, 2017, pp. 21–22.
- [15] M. H. Shahriar, Y. Xiao, P. Moriano, W. Lou, and Y. T. Hou, "CANShield: Signal-based intrusion detection for Controller Area Networks," 2022. [Online]. Available: <https://arxiv.org/abs/2205.01306>
- [16] M. E. Verma, R. A. Bridges, J. J. Sosnowski, S. C. Hollifield, and M. D. Iannaccone, "CAN-D: A modular four-step pipeline for comprehensively decoding Controller Area Network data," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 9685–9700, 2021.
- [17] C. Young, J. Svoboda, and J. Zambreno, "Towards reverse engineering Controller Area Network messages using machine learning," in *Proc. IEEE WF-IoT '20*, 2020, pp. 1–6.
- [18] M. D. Pesé, T. Stacer, C. A. Campos, E. Newberry, D. Chen, and K. G. Shin, "LibreCAN: Automated CAN message translator," in *Proc. ACM CCS '19*, ser. CCS '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2283–2300.
- [19] T. U. Kang, H. M. Song, S. Jeong, and H. K. Kim, "Automated reverse engineering and attack for CAN using OBD-II," in *Proc. IEEE VTC-Fall '18*, 2018, pp. 1–7.
- [20] M. Verma, R. Bridges, and S. Hollifield, "Actt: Automotive CAN tokenization and translation," in *Proc. CSCI '18*, 2018, pp. 278–283.
- [21] T. Huybrechts, Y. Vanommeslaeghe, D. Blontrock, G. Van Barel, and P. Hellinckx, "Automatic reverse engineering of CAN bus data using machine learning techniques," in *Proc. 3PGCIC '18*, F. Xhafa, S. Caballé, and L. Barolli, Eds. Cham: Springer International Publishing, 2018, pp. 751–761.
- [22] M. Markovitz and A. Wool, "Field classification, modeling and anomaly detection in unknown CAN bus networks," *Vehicular Communications*, vol. 9, pp. 43–52, 2017.
- [23] Comma.ai, "OpenDBC," <https://github.com/commaai/opendbc>, 2017.
- [24] M. E. Verma, M. D. Iannaccone, R. A. Bridges, S. C. Hollifield, P. Moriano, B. Kay, and F. L. Combs, "Addressing the lack of comparability & testing in CAN intrusion detection research: A comprehensive guide to CAN IDS data & introduction of the ROAD dataset," 2020. [Online]. Available: <https://arxiv.org/abs/2012.14600>
- [25] H. M. Song, H. R. Kim, and H. K. Kim, "Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network," in *Proc. ICOIN '16*, 2016, pp. 63–68.
- [26] A. Tomlinson, J. Bryans, S. A. Shaikh, and H. K. Kalutarage, "Detection of automotive CAN cyber-attacks by identifying packet timing anomalies in time windows," in *Proc. IEEE DSN-W '18*, 2018, pp. 231–238.
- [27] H. Olufowobi, C. Young, J. Zambreno, and G. Bloom, "SAIDuCANT: Specification-based automotive intrusion detection using Controller Area Network (CAN) timing," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1484–1494, 2020.
- [28] C. Young, H. Olufowobi, G. Bloom, and J. Zambreno, "Automotive intrusion detection based on constant CAN message frequencies across vehicle driving modes," in *Proc. AutoSec '19*, ser. AutoSec '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 9–14.
- [29] M.-J. Kang and J.-W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *PLOS ONE*, vol. 11, no. 6, pp. 1–17, 06 2016.
- [30] A. Taylor, S. Leblanc, and N. Japkowicz, "Anomaly detection in automobile control network data with long short-term memory networks," in *Proc. IEEE DSAA '16*, 2016, pp. 130–139.
- [31] S. Longari, D. H. Nova Valcarcel, M. Zago, M. Carminati, and S. Zanero, "CANnlo: An anomaly detection system based on LSTM autoencoders for Controller Area Network," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1913–1924, 2021.
- [32] K. Yang, X. Tang, S. Qiu, S. Jin, Z. Wei, and H. Wang, "Towards robust decision-making for autonomous driving on highway," *IEEE Transactions on Vehicular Technology*, vol. Early access, pp. 1–13, 2023.
- [33] "OBD-II PIDs (Wikipedia)," [https://en.wikipedia.org/wiki/OBD-II\\_PIDs](https://en.wikipedia.org/wiki/OBD-II_PIDs), accessed Jan. 1, 2023.
- [34] M. Müter and N. Asaj, "Entropy-based anomaly detection for in-vehicle networks," in *Proc. IEEE IV '11*, 2011, pp. 1110–1115.
- [35] M. Marchetti and D. Stabili, "Anomaly detection of CAN bus messages through analysis of ID sequences," in *Proc. IEEE IV '17*, 2017, pp. 1577–1583.
- [36] A. Taylor, N. Japkowicz, and S. Leblanc, "Frequency-based anomaly detection for the automotive CAN bus," in *Proc. WCICSS '15*, 2015, pp. 45–49.
- [37] M. Marchetti, D. Stabili, A. Guido, and M. Colajanni, "Evaluation of anomaly detection for in-vehicle networks through information-theoretic algorithms," in *Proc. 2nd IEEE RTSI '16*. IEEE, 2016, pp. 1–6.
- [38] D. Stabili, M. Marchetti, and M. Colajanni, "Detecting attacks to internal vehicle networks through hamming distance," in *Proc. AEIT '17*, 2017, pp. 1–6.
- [39] S. Katragadda, P. J. Darby, A. Roche, and R. Gottumukkala, "Detecting low-rate replay-based injection attacks on in-vehicle networks," *IEEE Access*, vol. 8, pp. 54 979–54 993, 2020.
- [40] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Vehicular Communications*, vol. 21, p. 100198, 2020.
- [41] H. M. Song and H. K. Kim, "Self-supervised anomaly detection for in-vehicle network using noised pseudo normal data," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1098–1108, 2021.
- [42] T.-N. Hoang and D. Kim, "Detecting in-vehicle intrusion via semi-supervised learning-based convolutional adversarial autoencoders," *Vehicular Communications*, vol. 38, p. 100520, 2022.



**Seonghoon Jeong** received a Ph.D. degree in information security from the School of Cybersecurity, Korea University, Seoul, Republic of Korea. He is currently a Postdoctoral Researcher with the Institute of Cybersecurity and Privacy, Korea University. His research has focused on in-vehicle network security, including intrusion detection systems for Controller Area Networks and automotive Ethernet.



**Sangho Lee** received the B.S. degree in electronic engineering from the Soongsil University in 2018, and the M.S. degree in information security from the School of Cybersecurity, Korea University in 2023. He is a security engineer at Samsung Research of Samsung Electronics since 2018. His research interests include data-driven security, user behavior analysis, and privacy.



**Hwejae Lee** received a B.S. degree in mechanical engineering from Kyung Hee University in 2020 and is a Ph.D. student in information security at the School of Cybersecurity, Korea University, Seoul, Republic of Korea. His research interests include vehicle security, data-driven security, intrusion detection, machine learning, and deep learning.



**Huy Kang Kim** received a Ph.D. degree in industrial and system engineering from the Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea. He founded A3 Security Consulting in 1999 and AI Spera, which is a data-driven cyber threat intelligence service company, in 2017. He is a Professor at the School of Cybersecurity, Korea University, Republic of Korea. His recent research has focused on intrusion detection in intelligent transportation systems and in-vehicle networks using machine-learning techniques.