2022 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung, November 8-9, 2022

# CANTool An In-Vehicle Network Data Analyzer

Md Rezanur Islam
Dept. of Smart Convergence Security
Soonchunhyang University
Asan, South Korea
arupreza@sch.ac.kr

Insu Oh
Dept. of Information Security Engineering
Soonchunhyang University
Asan, South Korea
catalyst32@sch.ac.kr

Kangbin Yim
Dept. of Information Security Engineering
Soonchunhyang University
Asan, South Korea
yim@sch.ac.kr

2022 International Conference on Information Technology Systems and Innovation (ICITSI) | 978-1-6654-6555-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITSI56531.2022.9970968

***Abstract*—In recent years, IoT devices have drawn attention to big data, complicating connectivity, and daily data processing. The automotive sector is no exception. The right way of vehicle data analysis is becoming essential every day for detecting internal errors, protecting against attackers, and connected vehicle concepts such as V2X. Some researchers use raw data to secure CAN, but that's not enough. On the other hand, deep learning is essential to secure autonomous driving and CAN, and data labeling is an obstacle. So, data analysis played an important role in data labeling. There are major flaws in data analysis, feature extraction, and data labeling for in-vehicle networks. Therefore, we proposed a CAN message analysis tool concept that can provide deep label analysis results and new features. There are many data analysis techniques these days, and we are trying to include suitable CAN message analysis techniques in our tool concept.**

***Keywords— CAN, Data Analysis, Feature Extraction.***

## I. Introduction

Automotive industry going through a huge transition point. In the past, vehicle concepts were mechanical, but modern vehicle components are a combination of mechanical and electronic components unique to hybrid vehicles and also electronic vehicles have also become popular because of their energy efficiency and less carbon emission. As a result, more and more modern vehicles are driven by wires, relying on constant communication from small computers called electronic control units (ECUs). Nearly ubiquitous in modern automobiles, controller area networks (CAN) facilitate the information exchange amongst ECUs by means of providing a common network with a standard protocol. On the other hand, in-vehicle networks have introduced many new features and are becoming more complex. According to this internal data diagnostic analysis and in-vehicle network, research work is carried out. For example, new concepts were introduced on vehicles like connected cars, autopilot, and the V2X (Communication between vehicle to everything) concept. While lightweight and reliable, the CAN standard has known security vulnerabilities as it lacks authentication, encryption, and other important security features [1]. Security researchers often use deep learning, so they need to label their data. Data labeling is directly dependent on data analysis. In-vehicle networks generate near about 2500 data per second and almost 70 to 100 ECUs consist in an in-vehicle network [2] and CAN data divided into nine segments. From this, it is easily understandable how difficult to analyze in-vehicle data. In this article, we are trying to discuss an efficient way to analyze CAN massage where statistic approach, frequency domain conversion approach and machine learning approach were implemented. Our previous study, used a statistical approach to distinguish the distribution patterns of CAN data for normal and various attacks [3].

## II. Characteristics of CAN

Data is transmitted in the controller area network via an ID, which contains a payload depending on the function. Fig. 1. Shows controller area network architecture. Four types of frames are available in CAN [4]. (1) Data Frame: This is the simplest frame used for payload switching. (2) Remote Frame: This part is used only to request payload transfers. When the ECU receives a remote frame, it responds immediately. (3) Error Frame: This framework defines and checks for possible errors. (4) Overload Frame: This frame is used to shift the start of the next message in case of congestion.

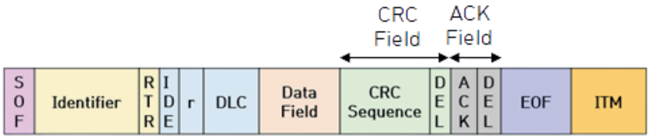


Fig. 1. CAN (Controller Area Network) Architecture

A CAN message payload has 8 bytes of hexadecimal, and each payload contains its own properties. The payload also has four parts [5]: (1) Constant, which is the unchanged value in the payload, (2) the main part, multi-values, and the main command that accompanies it. (3) counter is a circulation massage counter that starts over when the cycle is full. (4) Checkcodes. The last part of the payload has another error-checking mechanism developed in CAN.

CAN messages have no security mechanism installed, this is not entirely correct as the CRC is filed and the counter acts as a message authentication mechanism because the CRC field is generated from a fixed polynomial according to the payload. If the ECU gets a CRC mismatch, it means that the ECU did not accept the message, the counter counts the messages, and the range of each byte in the payload is 0-255. If the serial break is broken, that means ECU on the receiving side won't accept the message either, but it may still be vulnerable because these security mechanisms are only applicable to the receiver end. For this reason, proper data analysis is very important, not only for this reason but also for detecting errors in the network.

## III. Related Works

The in-vehicle network is a complex architecture, and it has different types of the scope of research and data analysis also depends on the goal of the research, for instance, inside communication response time, internal fault detection, intrusion detection. In-vehicle networks have many functions and message queues are cyclical, so operational tasks do not have the same priority [6]. For that reason, response time is a key point for the network. Most of the analysis about CAN is response time, in the paper [7] authors compare two different company ECUs according to response time and also they claim that pending massage data generation is not identical. Laufenberg & al. analyze different types of attack data by statistical method [8]. They collect all online available data and from analysis results they make assumptions. In recent days IDS (Intrusion Detection System) has become the most popular research on CAN and because of deep learning researchers approaches new data processing methods. Hossain & al. [9] used raw CAN traffic on their IDS but by using raw data need to design IDS for every cars individually. Some researchers convert data into image and some are converted data into frequency domain [10], [11] but still nobody introduce a complete template for CAN traffic analysis method. In this paper, we try to introduce a tool for CAN data analysis where different approaches can be implemented and two vehicles natural driving data presented as analysis example.

## IV. METHODOLOGY

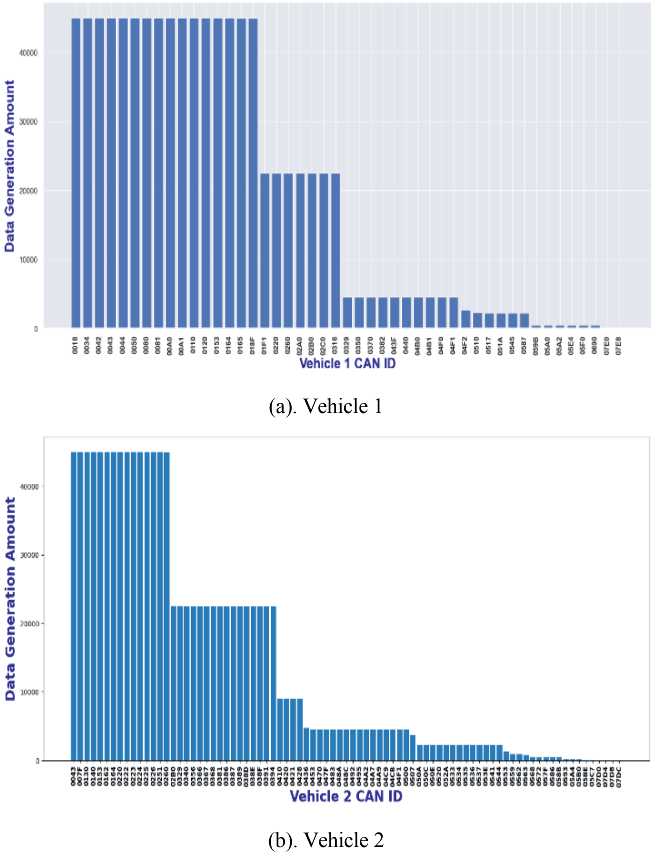
Python and pandas API opens a door for big data analysis in a convenient way and for mathematical operation, NumPy makes pandas more efficient. That platform gives the opportunity to analyze data deeply. Not only CAN traffic but also any kind of complex data like video and image data can be analyzed here by using different types of mathematical operations. For the analysis of our data, we used jupyter as an interactive computing platform. For frequency domain, FFT and wavelet here we used SciPy and pywavelets. Scikit learns machine learning unsupervised approach used by k-means and PCA (Principal Component Analysis). Last of all, for data visualization we used matplotlib and seaborn.

## V. CAN MASSAGE ANALYSIS

CAN message design according to the DBC file which is confidential by the manufacturer, and it differs from company to company. For that reason, the different car has different data even on the same company. In this paper, we take two cars' data as an example and will try to find similarities and significant differences from that dataset.

### A. Statistic Analysis

A natural driving, dataset with the same route and 450 seconds of data were used for both cars. For analysis, we used the Python Pandas platform [12], which is suitable for big data analysis. A total of 47 and 83 CAN IDs were found at this particular point in time and the data generation for each CAN ID is shown in Fig. 2. It's worth noting that one group produced the same amount. The main reason is that in a vehicle many functions such as RPM, speed, and braking are interdependent. Higher RPM means higher speed, and when a break is initiated, RPM decreases, and speed slows down. According to this, it can be said that for interdependent functions CAN ID generates data proportionally. In Fig. 2. (a) and (b) few groups of CAN ID generates in the same amount and there is a probability that those CAN IDs are interdependent.

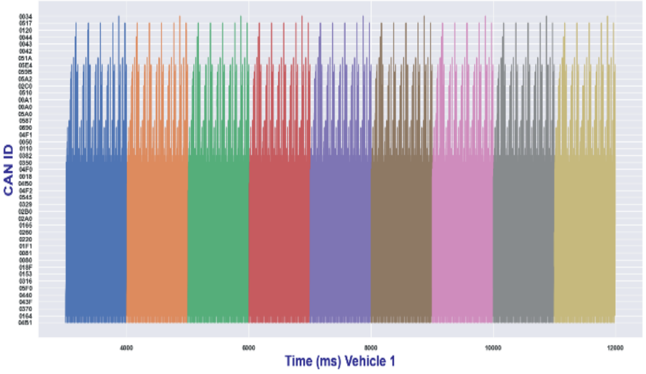


(a). Vehicle 1

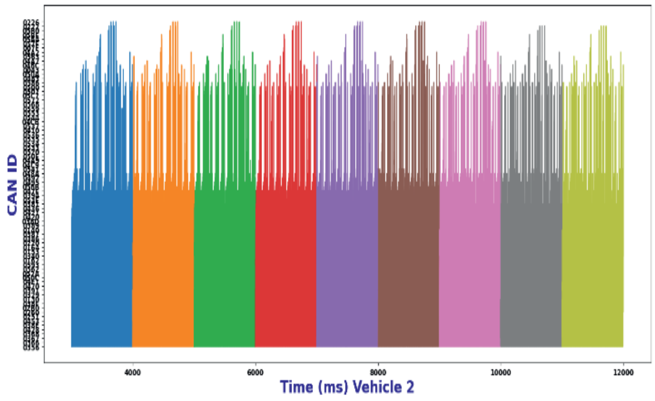
(b). Vehicle 2

Fig. 2. ID generation amount

In vehicles, networks are designed such a way that few functions are always in the operational mode for that reason a cyclic massage generation pattern follows with respect to time. In Fig. 3. (a) and (b) shows ID sequence for nine consecutive seconds. Every color represent one seconds and almost similar pattern seen for every second of data and another finding is for vehicle 1 for every second 45 CAN ID available and for vehicle 2 CAN ID range is 74 to 76.



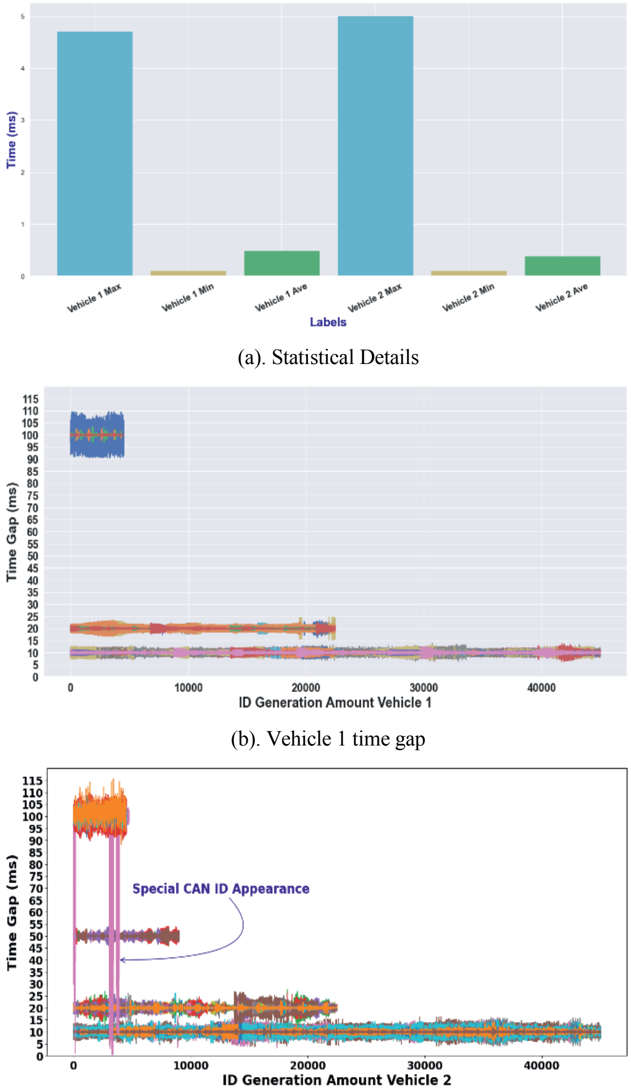
(a). Vehicle 1



(b). Vehicle 2

Fig. 3. ID sequence by time (milisecond)

Another point is the time interval between two messages. Under normal conditions, the time gap is almost fixed over the range of both the entire data set and ID wise filter. For the entire data set, Fig. 4. (a) shows the maximum, minimum, and average time gaps.



(a). Statistical Details

(b). Vehicle 1 time gap

(c). Vehicle 2 time gap

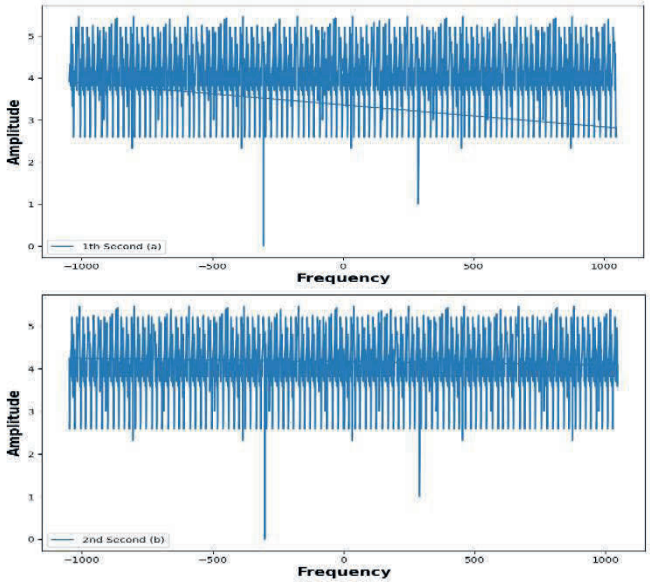
Fig. 4. Time gap analysis.

A CAN ID represents a vehicle function. Some IDs are infrequently used. In Fig. 4. (b). and Fig. 4. (c). We analyze ID wise time gap where we used the most frequently used CAN ID x-axis shows data generation amounts by an ID which means how many times an individual ID is generated over time and the y-axis is the time gap interval. According to the time interval per ID, vehicle 1 has 3 ID groups and vehicle 2 has 4 ID groups, and every group follows a specific time range, but all function is not used continuously in the vehicle for that reason the ID is generated when the user operated that particular function, as a result, some discontinuity has seen for some special ID Fig. 4. (c). From this feature extraction from vehicle, internal faults, and anomalies can be differentiated easily and it is possible to filter out time gap data for specific IDs from this group by using the CANTool concept.

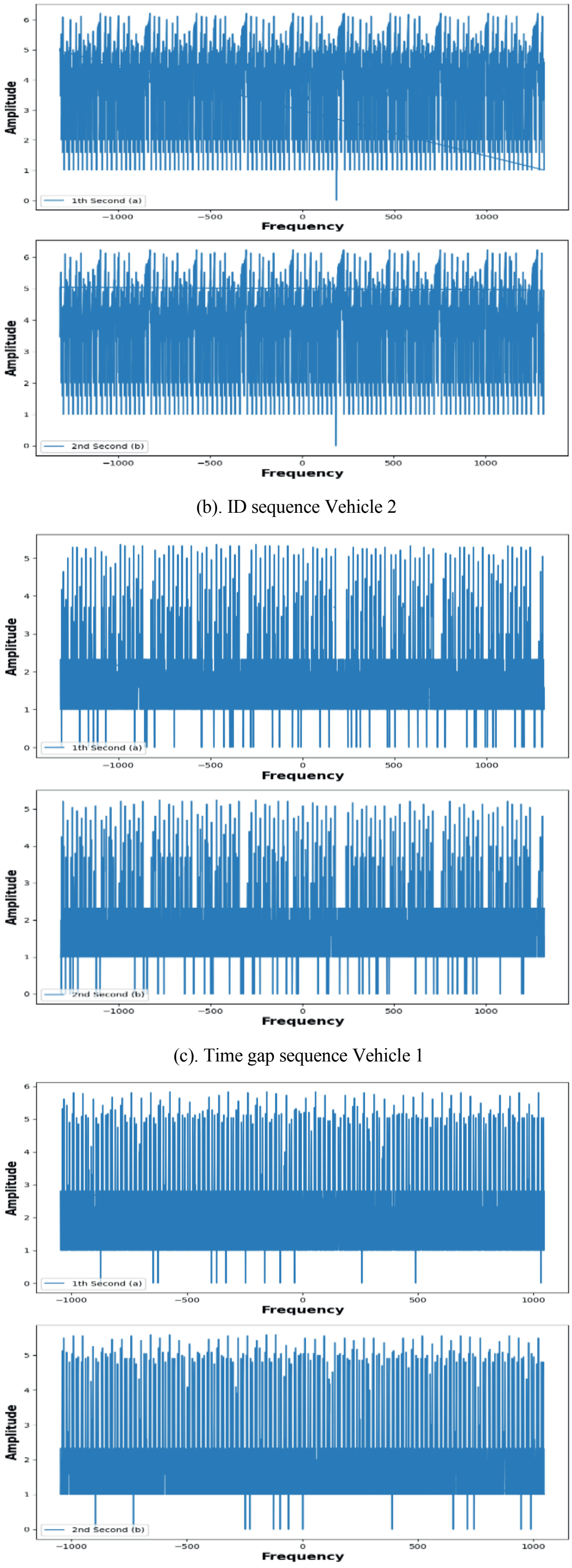
### *B. Frequency Domain Analysis*

Statistical analysis committed by time domain data, but another analysis become popular that is frequency domain analysis. On CAN massage frequency domain analysis mainly used for feature extraction for deep learning and the main goal for anomaly detection. Through time domain analysis a small change can be differentiated for that reason it become popular day by day. In first stage we applied FFT (Fast Fourier Transform) transformation and then for more advance label analysis we applied wavelet transformation analysis.

The Fourier transform can be used to convert a signal from the time domain to the frequency domain. The peaks in the frequency spectrum indicate the most prevalent frequencies of the signal. The larger and sharper the peak, the most common frequencies are present in the signal. The location (frequency value) and height (amplitude) of the frequency spectrum peaks can be used as inputs for the classifier.



(a). ID sequence Vehicle 1



(b). ID sequence Vehicle 2

(c). Time gap sequence Vehicle 1

(d). Time gap sequence Vehicle 2

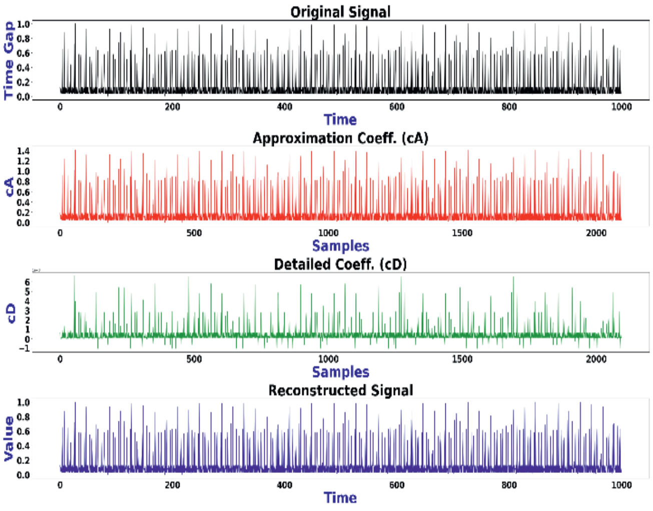
Fig. 5. FFT Analysis on CAN ID and Time gap Sequence.

From Fig. 5. the CAN ID sequence Fig. 5. (a). and Fig. 5. (b). and time gap sequence 5. (c). and 5. (d). are converted into FFT (Fast Fourier Transform), and it is clearly seen that in frequency analysis at a different time for both cars following a similar frequency pattern. We convert one-second data in an arbitrary way from 450 seconds. A large amplitude means that there is a large overlap between the two signals and that the signals that contain that particular frequency mostly exist. This is of course because the dot product is a measure of how much two vectors/signals overlap.

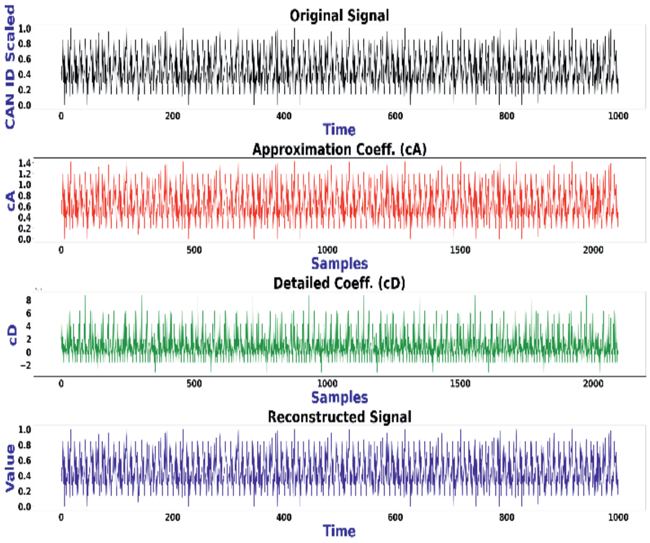
A characteristic of the Fourier transform is that it has high resolution in the frequency domain and no resolution in the time domain. That is, we know exactly what frequencies are present in the signal, but we do not know when those frequencies occurred. Another issue for FFT is that uncertainty principal which is theoretical limits of FFT [13]. For resolve this issue wavelet transform played an important role.

Wavelet transforms are high-resolution in both frequency and time domains. Not only do we know the frequencies in the signal, but we also know when those frequencies occurred. This is achieved by working at different scales. First look at the signal on a large scale/window and analyze the "large" features, then look at the signal on a small scale and analyze the small features [14]. This is the straightforward theory of wavelet transform. In wavelet transform two types of output produced first approximation coefficients vector and detailed coefficients vector, where original signal decomposes into low frequency band cA approximation coefficients and higher frequency band cD detailed coefficients.

We applied label 1 wavelet ‘bior6.8’ window filter. A lot of filter functions are available in the wavelet and label decide coefficient which means if the label increased then a more accurate value (which time which frequency generated) can be captured. In Fig. 6. Approximation coefficients represent the low pass of the signal and detailed coefficients are the wavelet coefficients. The reconstructed signal is the result of approximation and wavelet coefficients where the inverse transform is performed. In Fig. 6. (a). and Fig. 6. (b). shows CAN ID sequence and data generation time gap in wavelet form. For deep learning feature extraction from wavelet for CAN massage is a good medium [2].



(a). ID sequence



(b). Time gap sequence

Fig. 6. Wavelet Analysis vehicle 1

### C. *Machine Learning Approch for Data Analysis*

Unsupervised machine learning approach make the data analysis easier especially for CAN massage. For that reason, in our CAN analyzer, we introduced k-means and PCA clustering for classifying dataset into few groups. That means same type of data filter out by the algorithm. K-means is calculated by measuring the distance between each data point and its centroid, squaring that distance, and summing those squares across the cluster [15].

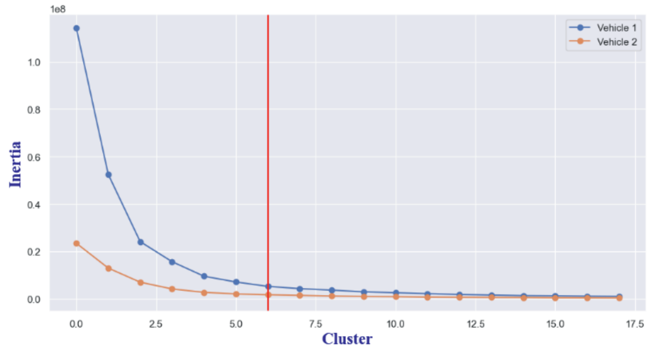
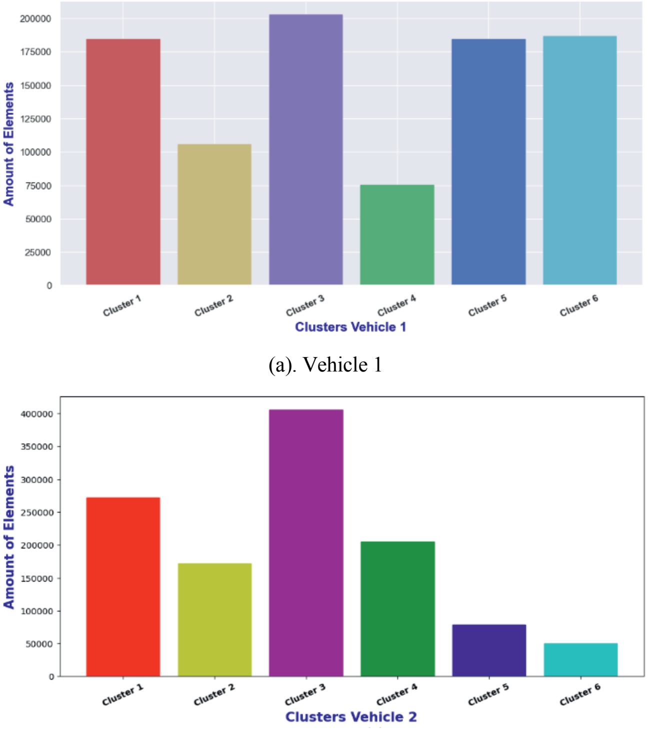


Fig. 7. K- Means Cluster Inertia

In Fig. 7. we input CAN ID sequence and time gap value for clustering in k-means and for both cars having six cluster of points. According to this cluster we can differentiate similarities and dissimilarities. Without CAN DBC file it is almost impossible to decode the CAN massage, but k-means open a research window for decode the CAN. By comparing functional response data on k-means it is partially possible to decode the CAN.

This section describes clusters where each cluster corresponds to an inertia property. Fig. 8. shows the amount of data available in each cluster. Fig. 8. (a). vehicle 1 has less data generation volume and CAN ID volume than Fig. 8. (b). vehicle 2, vehicle 1 is an older model and vehicle 2 has been updated with new features. From this point on, data generation is proportional to the CAN ID. Fig. 8. (a). and Fig. 8. (b). y-axis shows the data distribution amount. Each cluster has a centroid point that can be analyzed in greater depth depending on the cluster.



(a). Vehicle 1

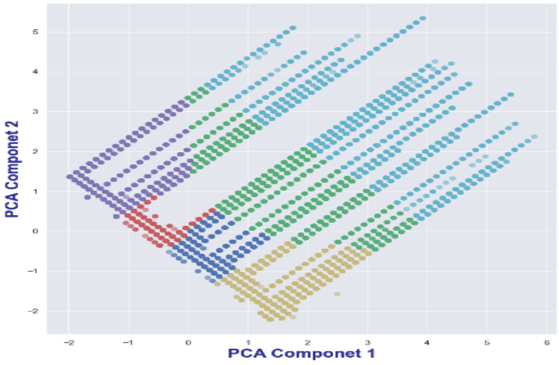
(b). Vehicle 2

Fig. 8. Cluster data distribution details

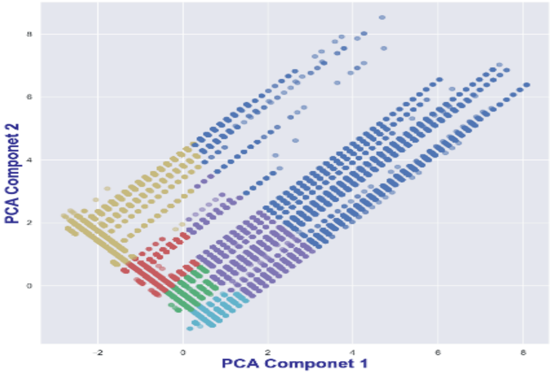
For cluster-by-cluster analysis, here we present a demo where each cluster time-gap sequence is displayed. Table 1 details the time gap for each cluster, and all clusters have the same minimum time gap of 0.09 ms. Also, each cluster has a range of upper and lower bounds. Fig 9. Another clustering method applied PCA (Principal Component Analysis) which is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables. Fig 9. (a) and Fig 9. (b) vehicle 1 and vehicle 2 cluster distribution and each color represent a cluster and an identical similarity overview found here.

TABLE I. CLUSTER DETAILS

| Cluster Name | Maximum | Minimum | Average |
|---|---|---|---|
| Cluster 1 | 5.30 ms | 0.09 ms | 1.65 ms |
| Cluster 2 | 19.40 ms | 0.09 ms | 2.61 ms |
| Cluster 3 | 8.79 ms | 0.09 ms | 1.11 ms |
| Cluster 4 | 10.60 ms | 0.09 ms | 2.19 ms |
| Cluster 5 | 25.19 ms | 0.09 ms | 5.70 ms |
| Cluster 6 | 52.29 ms | 0.09 ms | 9.03 ms |



(a). Vehicle 1



(b). Vehicle 2

Fig. 9. PCA cluster comparison

## VI. RESULT & DISCUSSION

In statistical analysis, Fig. 2. ID-wise data distribution represented where co-related ID can be found easily by comparing similar data generated by different IDs. CAN data generation is a continuous process, so data generation and time have a relation and Fig. 3. Time-wise analysis approach represented and in a specific time interval pattern repetition might be a feature extraction for deep learning. According to the time gap, every ID follows a pattern. From this pattern, co-related CAN ID can be found, and special function ID also might be notified. From Fig. 2. ID-wise data generation, Fig. 3. and Fig. 4. time sequence, time gap sequence, and finally cluster analysis Fig. 8. And Fig. 9. from those combined analyses opens an opportunity for partial CAN message translation. Time gap sequence has already been established as an intrusion detection feature. Frequency domain analysis Fig. 5. and Fig. 6. are mainly used for finding small changes and for patterns. For fault detection and label data, frequency domain analysis will be an efficient technique. Unsupervised clustering Fig. 8. and Fig. 9. the main goal is to filter out the same pattern data. After that, every cluster can be analyzed statistically and frequency conversion.

Here we represent three types of techniques for data analysis. It is not only applicable to CAN traffic but also to any kind of data. By using those techniques data labeling also becomes easier, especially for big arbitrary data, and python libraries provide interactive graph results. As a result, in big data implementing multidimensional techniques makes analysis more convenient and fruitful.

## VII. CONCLUSION

This article focuses on proper methods for analyzing CAN. For confidentiality reasons, CAN messages remain a mystery, making it difficult to standardize the analysis. Here we only analyze on CAN ID sequence and time gap same way we can analyze on payload and CRC field and the payload can be divided into each byte called PID (parameter ID). CAN messages contain many parts and are difficult to parse. To minimize this complexity, we divided the analysis process into three parts. We compared two different vehicles and found some similar features. But over time, the process of data generation in in-vehicle networks changed as car companies introduced new features. Our future roadmaps, we will use these approaches to figure out how to decode CAN messages.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1A4A2001810) and the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-01343, Regional strategic industry convergence security core talent training business).

## REFERENCES

[1] M. E. Verma et al., “Addressing the Lack of Comparability & Testing in CAN Intrusion Detection Research: A Comprehensive Guide to CAN IDS Data & Introduction of the ROAD Dataset,” IEEE Trans. Veh. Technol., vol. XX, p. 1, Dec. 2020, doi: 10.48550/arxiv.2012.14600.

[2] M. R. Islam, I. Oh, M. Batzorig, M. Kim, and K. Yim, “Wavelet Transform Based PID Sequence Analysis for IDS on CAN Protocol,” vol. 2, 2022, pp. 85–96.

[3] M. R. Islam, I. Oh, M. Batzorig, S. Kim, and K. Yim, “A Concept of IDS for CAN Protocol Based on Statics Theory,” in Lecture Notes in Networks and Systems, 2022, pp. 294–302.

[4] [4] H. Lee, S. H. Jeong, and H. K. Kim, “OTIDS: A Novel Intrusion Detection System for In-vehicle Network by Using Remote Frame,” in 2017 15th Annual Conference on Privacy, Security and Trust (PST), Aug. 2017, pp. 57–5709, doi: 10.1109/PST.2017.00017.

[5] M. Markovitz and A. Wool, “Field classification, modeling and anomaly detection in unknown CAN bus networks,” Veh. Commun., vol. 9, pp. 43–52, Jul. 2017, doi: 10.1016/j.vehcom.2017.02.005.

[6] H. Zeng, M. Di Natale, P. Giusto, and A. Sangiovanni-Vincentelli, “Statistical analysis of controller area network message response times,” Proc. - 2009 IEEE Int. Symp. Ind. Embed. Syst. SIES 2009, pp. 1–10, 2009, doi: 10.1109/SIES.2009.5196185.

[7] K. W. Tindell, H. Hansson, and A. J. Wellings, “Analysing real-time communications: Controller area network (CAN),” Proc. - Real-Time Syst. Symp., pp. 259–263, 1994, doi: 10.1109/REAL.1994.342710.

[8] J. Laufenberg, T. Kropf, and O. Bringmann, “Static Analysis of Controller Area Network Communication for Attack Detection,” Eur. J. Secur. Res., vol. 6, no. 2, pp. 171–187, 2021, doi: 10.1007/s41125-021-00077-1.

[9] 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, 2020, doi: 10.1109/ACCESS.2020.3029307.

[10] E. Seo, H. M. Song, and H. K. Kim, “GIDS: GAN based Intrusion Detection System for In-Vehicle Network,” 2018, doi: 10.1109/PST.2018.8514157.

[11] M. Bozdal, M. Samie, and I. K. Jennions, “WINDS: A Wavelet-Based Intrusion Detection System for Controller Area Network (CAN),” IEEE Access, vol. 9, pp. 58621–58633, 2021, doi: 10.1109/ACCESS.2021.3073057.

[12] “pandas - Python Data Analysis Library.” https://pandas.pydata.org/ (accessed Sep. 09, 2022).

[13] S. Parsons, A. M. Boonman, and M. K. Obrist, “ADVANTAGES AND DISADVANTAGES OF TECHNIQUES FOR TRANSFORMING AND ANALYZING CHIROPTERAN ECHOLOCATION CALLS,” J. Mammal., vol. 81, no. 4, pp. 927–938, Nov. 2000, doi: 10.1644/1545-1542(2000)081<0927:AADOTF>2.0.CO;2.

[14] S. L. Brunton and J. N. Kutz, Data-Driven Science and Engineering. Cambridge University Press, 2019.

[15] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means Algorithm: A Comprehensive Survey and Performance Evaluation,” Electron. 2020, Vol. 9, Page 1295, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/ELECTRONICS9081295.