Heliyon 10 (2024) e25000

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The typical AV accident scenarios in the urban area obtained by clustering and association rule mining of real-world accident reports

Hojun Lee [a,1], Minhee Kang [b,1], Keeyeon Hwang [b,**], Young Yoon [c,*]

[a] SpaceInsight Co., Ltd., Seoul, 07788, Republic of Korea
[b] Department of Electrical Engineering, Korean Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea
[c] Department of Computer Engineering, Hongik University, Seoul, 04066, Republic of Korea

ARTICLE INFO

Keywords:
Automated vehicles
Safety evaluation
Typical accident scenarios
Real-world data

ABSTRACT

Automated Vehicles (AVs) based on a collection of advanced technologies such as big data and artificial intelligence have opened an opportunity to reduce traffic accidents caused by human drivers. Nevertheless, traffic accidents of AVs continue to occur, which raises safety and reliability concerns about AVs. AVs are particularly vulnerable to accidents on urban roads than on highways due to various dynamic objects and more complex infrastructure. Several studies proposed a scenario-based approach of experimenting with the response of AVs to specific situations as a way to test their safety. Reliable and concrete scenarios are necessary to test AV safety under critical conditions accurately. This study aims to derive a typical accident scenario for evaluating the safety of AVs, specifically in urban areas, by analysing collisions reported by the DMV of California, USA. We applied a hierarchical clustering method to find groups of similar reports and then executed association rule mining on each cluster to correlate between accident factors and collision types. We combined statistically significant association rules to constitute a total of 14 scenarios that are described according to an adapted PEGASUS framework. The newly obtained scenarios exhibit significantly different accident patterns than the typical Human-driven Vehicles (HVs) in urban areas reported by National Highway Traffic Safety Administration. Our discovery urges AV safety to be tested reliably under scenarios more relevant than the existing HV accident scenarios.

1. Introduction

Automated Vehicles (AVs) developed with state-of-the-art artificial intelligence (AI) and big data technologies have brought positive effects such as improved mobility and reduced social costs through AVs' contribution to the decrease of traffic accidents [1]. SAE [2] classified AVs into six levels (lv.0 - lv.5). At a higher level, AVs are more capable of coping with various situations with less human intervention. Some AVs at lv.3 are on the market, and several companies such as GM, Google, and Nuro are conducting test runs

* Corresponding author.
** Corresponding author.
E-mail addresses: higeree@naver.com (H. Lee), ministop@kaist.ac.kr (M. Kang), hwangkeeyeon@kaist.ac.kr (K. Hwang), young.yoon@hongik.ac.kr (Y. Yoon).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e25000
Received 4 April 2023; Received in revised form 22 November 2023; Accepted 18 January 2024
Available online 19 January 2024

of lv.4 AVs that require driver intervention only under certain critical conditions. Despite recent advancements, AVs are not entirely free from accidents. It is reported that accidents involving AVs constantly have occurred for various reasons, even under ideal conditions [3,4]. Furthermore, the recent fatal pedestrian fatality involving AVs is causing fears and hampering the successful commercialization of AVs.

The research community has tried to understand AV accidents by referring to the accident data collected from Human-driven Vehicles (HVs) [5]. However, AVs could exhibit different accident patterns compared to HVs. Instead, a few researchers looked into AVs accident data made publicly available on the web by the state of California [6–11]. The Department of Motor Vehicles (DMV) of California mandates the submission of collision reports in case of accidents during AV test driving. However, most previous studies have relied on manual analysis of accident characteristics. With such approaches, AV test scenarios that are necessary for evaluating the safety functions of AVs cannot be efficiently generated [12–15]. In addition, the manual analysis by humans can yield subjective and biased conclusions.

We are keen to employ data-driven methods based on clustering and association rule mining algorithms for more efficient and objective analysis. This paper focuses mainly on urban-area accidents in which AVs are prone to get involved due to the non-trivial interactions in an open system with complex infrastructure and pedestrians besides vehicles [16,17]. Urban area accidents are more frequent and follow patterns that differ from highway accidents. Accidents on the highways can be relatively much more fatal. However, AV encountering non-vehicular objects on highways is less likely [18]. Given the result of the data-driven analysis of DMV collision reports, we compose accident scenarios according to the framework proposed by the PEGASUS project.

The contribution of our research work can be summarized as follows:

(1) We have devised data-driven analytical tools to extract statistically significant accident patterns from 165 DMV collision reports involving AVs.
(2) We have generated 313,748 correlations (association rules) between 14 accident factors and collision types.
(3) We implemented a novel method to combine any two association rules to constitute an AV accident scenario. As a result, we have derived 14 unique AV accident scenarios that are specified according to an adapted PEGASUS scenario description framework.
(4) We have confirmed a significant difference between the AV and HV accident patterns by comparing our newly generated scenarios with the urban-area HV pre-crash reports by NHTSA (National Highway Traffic Safety Administration). Such discovery justifies our work of analyzing AV accident patterns to derive safety test scenarios more relevant for AVs.

In summary, we utilized AV data and employed data mining techniques such as clustering and association rule mining to ensure both the reliability and the concreteness of AV accident scenarios. These scenarios can effectively prevent AV accidents.

The remainder of this paper is structured as follows: In Section 2, we examine previous studies regarding data utilization and scenario derivation for AVs; In Section 3, we introduce the format and semantics of DMV collision reports, the data preprocessing procedure and the data analysis methodology; In Section 4, we evaluate the analysis results; In Section 5, we present a few cases of urban accident scenarios we derived based on the results from Section 4; Finally, we conclude in Section 6.

2. Related works

This section reviews studies that analyze traffic accidents in HVs and AVs and derive accident scenarios for both vehicles.

Shanthi and Ramani (2012) [19] predicted the injury severity in traffic accidents using various classification algorithms, and they derived factors affecting the injury severity. With an accuracy of 99.73 %, they confirmed that the vehicle's collision type, seat position, age group, and drug significantly affected the injury severity. Taamneh et al. (2016) [20] used several classification methods, such as decision trees and Naive Bayes, to select age, gender, and collision type as major accident factors related to injury severity in traffic accidents. The study by Taamneh et al. revealed that 18-30-year-olds were vulnerable to traffic accidents, and driver injuries occurred more frequently than pedestrians and passengers. Muhammad et al. (2017) [21] analyzed the cause of the accident using an ID3 decision tree based on accident data on the Kano-Wudil Highway. As a result, they found that incorrect passing, loss of control of the vehicle, tire puncture, and brake failure were the main causes of the accident. In addition, Bahiru et al. (2018) [22] conducted a study to classify and predict accident severity, and they found that weather conditions, traffic lanes, and accident times were essential factors influencing accident severity. Janani and Devi (2016) [23] and Li et al. (2017) [24] identified the characteristics of traffic accidents using classification, clustering, and association rules. Li et al. identified that roadway surface, weather, and light conditions did not affect the fatality rate of traffic accidents, while driving under the influence or the collision type substantially affected the fatality rate. Boggs et al. (2020) [25] confirmed that most accidents (61 %) were rear-end collisions, and 13.3 % were injury accidents through text analysis of the DMV collision report. Furthermore, through the setting of random parameters in the Bayesian analysis, they found that the correlation between disengagement of the autonomous driving system and rear-end collision was significantly high, and the possibility of AV rear-end collisions were quite low near public/private schools. Lee et al. (2023) [26] analyzed the AV accident using a DMV collision report, and they identified the correlation between pre-crash conditions, AV driving modes, crash types, and crash outcomes. This study argued that in autonomous driving mode, AV should pay attention to the longitudinal distance from the front or back vehicle and that road and infrastructure functions such as intersections, ramps, and slip lanes need to be improved.

Several studies used clustering methods to extract characteristics from data to determine the cause of traffic accidents. In addition, association rules were mined to identify the relationship between variables in the data. DrissiTouzani et al. (2020) [27] classified accidents into 10 clusters through the K-means clustering algorithm, and they confirmed that most traffic accidents occurred during the day, and the type of collision, initial shock point, and movement of the vehicle were essential factors in traffic accidents. Lin et al. (2014) [28] and Kumar and Toshniwal (2016) [29] derived significant and robust association rules using both clustering and association analysis, and they emphasized that clustering helped to narrow down the characteristics and patterns before the association analysis. Also, Kong et al. (2022) [30] performed cluster correspondence analysis using dimension reduction based on near-crash data. As a result, they derived six clusters with four types: near-crash with adjunct vehicles; near-crash with the following or leading vehicles; and near-crash with objects on the road. This study also found characteristics of the near-crash such as slow or stopped, rapid deceleration or stop of the leading vehicle. Wang et al. (2022) [31] derived six clusters through k-medoid clustering using two-wheeled vehicle crash data in China. Based on this analysis, they presented functional, logical, and concrete scenarios specifically for AV test.

Studies on accident prevention technology development and scenario generation for vehicle safety evaluation have been conducted based on real-data analysis [32–40]. Nitsche et al. (2017) [32] derived 34 crash scenarios using clustering and association analysis and extracted 12 scenarios of accidents with a high risk of injury. Sui et al. (2019) [33] derived car-two-wheeler test scenarios with clustering, and they derived six scenarios. Yuan et al. (2020) [34] derived a high-risk scenario using ANN based on HV accident data. They set time, weather, road type, and speed limitations as input, and five high-risk scenarios were derived considering the accident frequency and probability distribution. Kong et al. (2021) [35] analyzed the pattern of near-crash events through association rule mining between factors other than secondary tasks. They used VCC50 Elite data from 50 connected vehicles with adaptive cruise control and lane-keeping assistance functions. In addition, they utilized the apriori algorithm and performed the association rule mining by setting the presence or absence of a secondary task to the consequent. This paper confirmed that the rapid deceleration or stop of the leading vehicle, regardless of the secondary task, is highly related to the near crash. In the case of a near-crash without a secondary task, the rapid deceleration or stop of the leading vehicle is positively associated with lane change and sideswipe accidents. Pan et al. (2021) [36] and Tan et al. (2021) [37] derived accident test scenarios by clustering analysis to develop AEB/FCW technology to prevent accidents practically. Essenturk et al. (2022) [38] derived traffic accident patterns through ROCK (Robust Clustering with Links) and market basket analysis using UK's STATS19 database. ROCK was performed on 26 clusters (derived from clustering) to create seven clusters, and an AV test scenario was presented. This study significantly contributes to the derivation of AV test scenarios by employing clustering, ROCK analysis, and market basket analysis. However, it is important to acknowledge certain limitations associated with the use of HV data in this research. Kang et al. (2022a) [39] used a vision transformer to detect critical situations involving AVs. The vision transformer caught critical situations at an F1 score of 94 %. Given the interpretation of the result returned by the vision transformer, Kang et al. followed the PEGASUS framework to derive accident scenarios under which AV safety functions can be tested. Liu et al. (2021) [5] extracted accident characteristics of HVs data based on the NHTSA pre-crash scenario and extracted aspects of AVs accidents. Liu et al. also developed a scenario for evaluating the safety of AVs by comparing the characteristics of two vehicles. However, this study pointed out a limitation to applying HVs data-based accident scenarios to AVs due to the different characteristics of the two vehicles, such as differences in perception-response time (PRT) between human and system drivers. Accordingly, it is necessary to derive a reliable scenario using AV data to secure the practical safety of AVs. Similarly, Novat et al. (2023) [40] conducted a study comparing the accident characteristics of AV and conventional vehicles through the Bayesian network using CA BMV (California Bureau of Motor Vehicles) data. This study emphasized that AV and conventional vehicles have different collision patterns. For instance, AVs had more rear-end collisions than conventional vehicles, and sideswipe/broadside and other collision types were less likely to occur. Their study does not deal with the generation of accident scenarios.

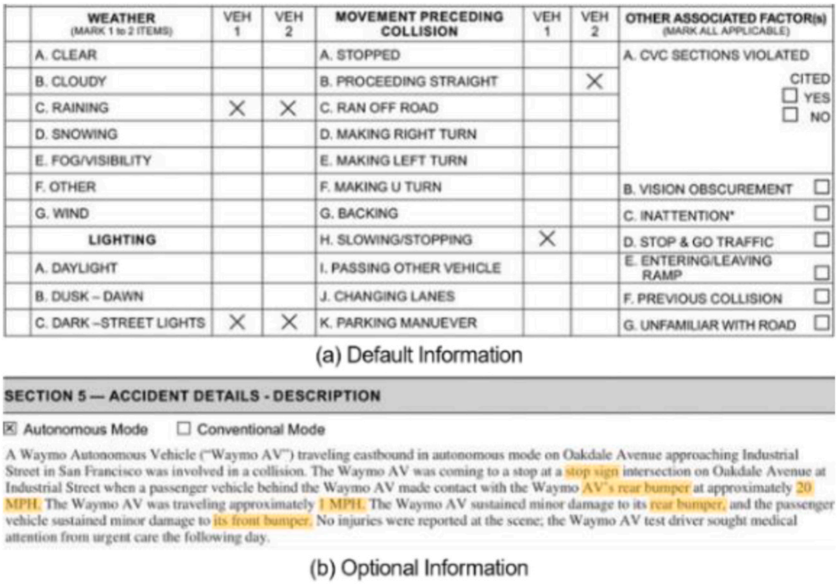
Although accident analysis and scenario generation studies based on AVs data are required, the manufacturer's confidential information restricts access to AV driving data. On the other hand, the DMV in California, USA, is obliged to submit accident reports during AV tests. The report is made available on the DMV website, and the data has been used in various ways in recent studies on AVs accident analysis. Most studies affirmed that AVs' representative accident collision type was a rear collision [9,41–44]. Leilabadi and Schmidt (2019) [9] found a strong correlation between the road surface and the accident's severity and that the type of accidents appeared differently depending on the driving mode (automated mode and conventional mode). Das et al. (2020) [41] identified six significant collision patterns, including left/right turn and proceeding straight before the collision. Torres et al. (2021) [42] analyzed the DMV report using multimedia logic regression, and they affirmed that the movement before the collision of AVs had a significant influence on the collision type of the accident. Ashraf et al. (2021) [43] derived rules for AV accidents using a decision tree and association analysis. Most accidents occurred when the AVs stopped in the automated mode at the intersection. Also, Ashraf et al. found that an accident frequently occurred when HVs passed the AVs or two vehicles turned. Kang et al. (2022b) [44] conducted random forest analysis by constructing a DMV collision report and presented more than 100 autonomous vehicle accident scenarios based on the random forest results.

A few research works tested the response of AVs to a specific situation [12–15]. For such tests, scenario derivation studies have been conducted. Stark et al. (2020) [45] raised the need to generate scenarios using accident data exclusively for AVs. Alambeigi et al. (2020) [11] derived situations that should be noted in AV accidents through topic modeling analysis. This study suggested that AV manufacturers need to check the following incidents in-depth: (1) manual transitions; (2) side collisions; and (3) rear-end collisions due to left/right turns in cross-section. Song et al. (2021) [18] proposed seven typical crash scenarios by analyzing the accident characteristics of AVs through clustering analysis and sequence analysis based on DMV collision report and DMV disengagement report.

These previous research works came short in reliability and concreteness. Nitsche et al. (2017) [32] secured the concreteness of the AVs scenario through clustering analysis and association analysis but failed to secure reliability by using data from HV accidents that may exhibit different patterns than AV accidents. In addition, Song et al. (2021) [18] composed AVs scenario based on the data from AV accidents. However, their proposed scenarios were vaguely written with limited factors such as vehicle movement and collision type.

Table 1
A comparison of our methods with previous studies.

Studies	Scenario Derivation	AV Accident Data Usage	Clustering	Association Rule Mining
.Alambeigi et al. (2020) [11]	X	O	X	X
Liu et al. (2021) [5]	O	O	X	X
Sui et al. (2019) [33]	O	X	O	X
Nitsche et al. (2017) [32]	O	X	O	O
Song et al. (2021) [18]	O	O	O	X
Our methods	O	O	O	O



WEATHER (MARK 1 to 2 ITEMS)	VEH 1	VEH 2	MOVEMENT PRECEDING COLLISION	VEH 1	VEH 2	OTHER ASSOCIATED FACTOR(s) (MARK ALL APPLICABLE)
A. CLEAR			A. STOPPED			A. CVC SECTIONS VIOLATED
B. CLOUDY			B. PROCEEDING STRAIGHT		X	CITED ☐ YES ☐ NO
C. RAINING	X	X	C. RAN OFF ROAD			
D. SNOWING			D. MAKING RIGHT TURN			
E. FOG/VISIBILITY			E. MAKING LEFT TURN			
F. OTHER			F. MAKING U TURN			B. VISION OBSCUREMENT ☐
G. WIND			G. BACKING			C. INATTENTION* ☐
LIGHTING			H. SLOWING/STOPPING	X		D. STOP & GO TRAFFIC ☐
A. DAYLIGHT			I. PASSING OTHER VEHICLE			E. ENTERING/LEAVING RAMP ☐
B. DUSK – DAWN			J. CHANGING LANES			F. PREVIOUS COLLISION ☐
C. DARK –STREET LIGHTS	X	X	K. PARKING MANUEVER			G. UNFAMILIAR WITH ROAD ☐

(a) Default Information

SECTION 5 — ACCIDENT DETAILS - DESCRIPTION

☒ Autonomous Mode ☐ Conventional Mode

A Waymo Autonomous Vehicle ("Waymo AV") traveling eastbound in autonomous mode on Oakdale Avenue approaching Industrial Street in San Francisco was involved in a collision. The Waymo AV was coming to a stop at a stop sign intersection on Oakdale Avenue at Industrial Street when a passenger vehicle behind the Waymo AV made contact with the Waymo AV's rear bumper at approximately 20 MPH. The Waymo AV was traveling approximately 1 MPH. The Waymo AV sustained minor damage to its rear bumper, and the passenger vehicle sustained minor damage to its front bumper. No injuries were reported at the scene; the Waymo AV test driver sought medical attention from urgent care the following day.

(b) Optional Information

Fig. 1. Default & optional value of DMV collision report.

Scenario-based testing is an important approach as a method of assessing AV safety. The scenario requires reliability, and this reliability is based on data. Since AV and HV accident patterns are different [5,40], scenarios created based on AV data should be reliable. In addition, the scenario consisting of various situations that threaten the safety of AV must clearly present various factors for AV testing. It is concreteness, so the scenario needs to secure concreteness to test a range of situations. Therefore, we found that most AV accident scenarios derived from most studies lack reliability or concreteness. To address these critical shortcomings, our research possesses the following characteristics as shown in Table 1. We derived AV accident scenarios based on clustering and association rule mining to enhance the concreteness of the scenarios. Notably, our research aims to secure the reliability of the AV accident scenario by utilizing real-world AV accident reports. Also, we have adapted the standard scenario format, such as the PEGASUS project.

3. Accident data analysis methodology

In this section, we introduce analysis methods based on the data extracted from the DMV reports.

3.1. Data collection

We constructed an AV accident database with structured data from collision reports archived at DMV in California, USA. Currently, California permits the manufacturers such as Waymo and GM to test drive their AVs on real roads. If an accident occurs during test driving, the manufacturers must file a report and make it publicly available on the DMV website [46]. The report specifies the manufacturer of AVs, driving mode, road conditions, the type of collision, and the movement preceding the collision of vehicles related to the accident. AVs and HVs involved in the collision report are denoted as VEH 1 and VEH 2, respectively, as shown in Fig. 1. Since the revision in 2018, DMV additionally specifies whether the accident occurred while driving in an automated or manual mode. Out of 370 reports since 2018, we narrowed it down to 165 collision reports that occurred while AVs were in an automated mode.

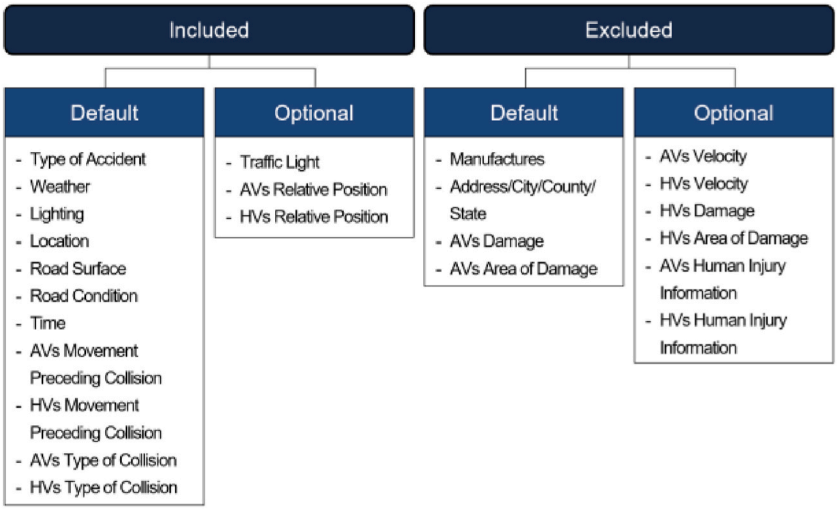


Fig. 2. Selection from the DMV collision report.

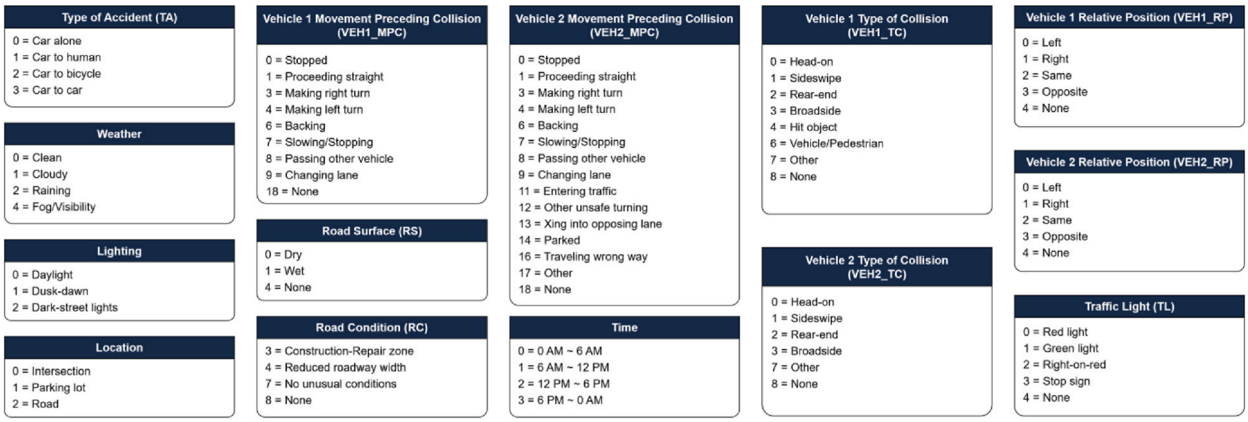


Fig. 3. Encoding values of accident factors.

3.2. Accident factors and encoding

The format of a DMV collision report is shown in Fig. 1. From the report, we identified accident factors as follows.

The most critical factor is the collision type, expressed as a 2-tuple. The first and second elements of the tuple specify AV and HV body parts involved in an accident or their motions at the moment of collision, respectively.

For instance, (Sideswipe, Sideswipe) indicates that an AV and an HV swiped each other on the side. Another example, such as (Hit object, None), means that an AV hit a non-vehicular object. AV and HV movements and their relative positions in terms of lanes are also expressed at description as shown in Fig. 1-(b).

From the optional written description section of the report, we first extracted additional fields to consider, as shown in Fig. 2. We excluded manufacturer, damage, and human injury information since they are not the cause of an accident. Over 80 % of the velocity information was missing from the reports. Since we could not ensure the statistical significance, we had to ignore the velocity fields despite their importance in inferring the correlation with an accident if enough information was provided. We did not use the precise address of the accident location. Instead, we classified the address to a location type, such as intersections, roads, and parking lots. The location types are added to the list of accident factors.

We compiled the final list of 14 accident factors, as shown in Fig. 2. The encoded values for each factor are specified in Fig. 3. The value N/A refers to the information that was not available in the report.

3.3. Preliminary statistics of DMV collisions

We discovered a few preliminary statistical characteristics of the DMV reports as follows.

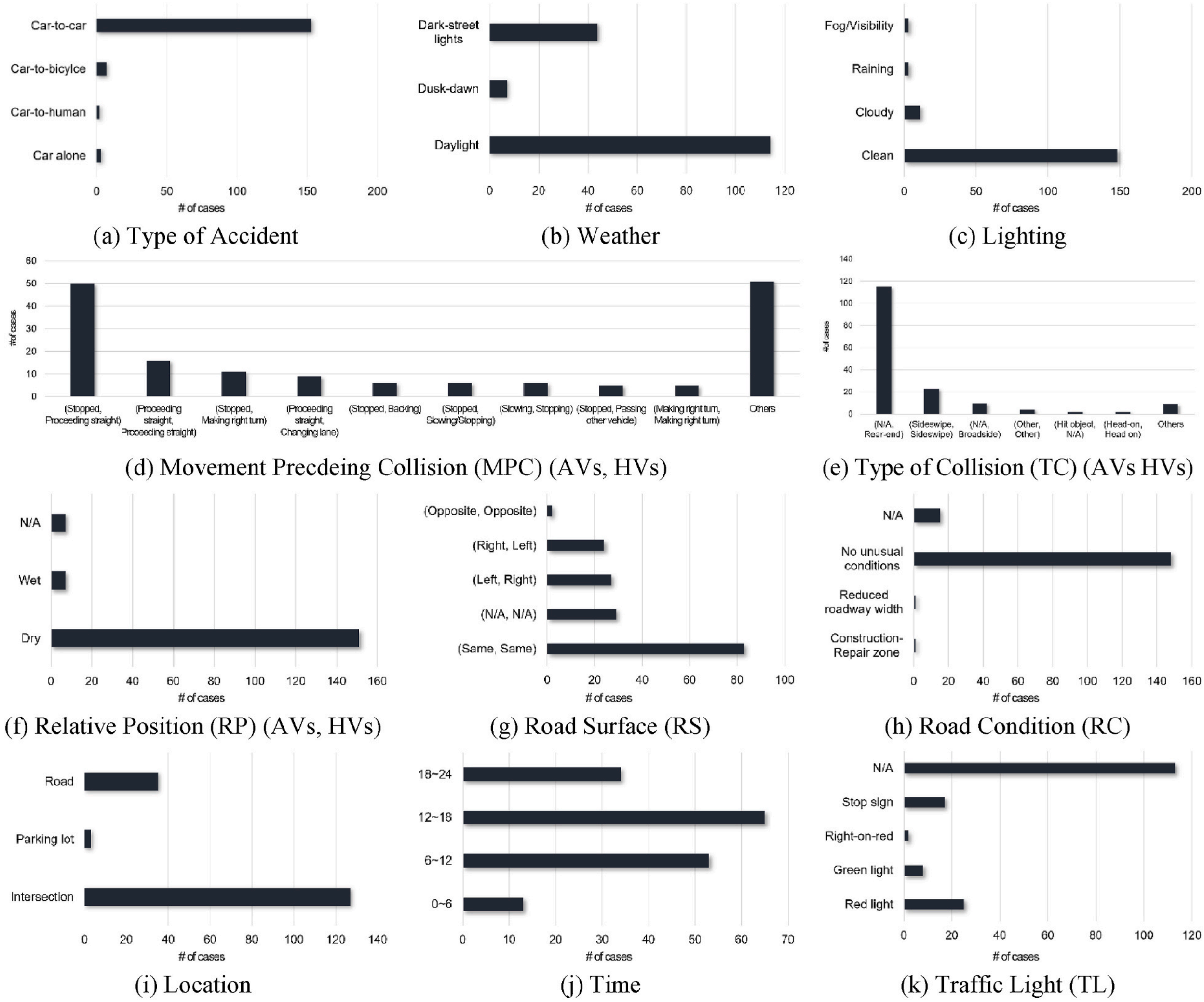


Fig. 4. Result of preliminary statistical analysis.

1. We found that 93 % of the accident were car-to-car collisions (Fig. 4-(a)).
2. 90 % of the accidents occurred in clean weather (Fig. 4-(b)).
3. 69 % of the accidents occurred in daylight (Fig. 4-(c)).
4. In 30 % of the accidents, AVs stopped while HVs proceeded straight prior to the collision, i.e., (Stopped, Proceeding straight) (Fig. 4-(d)).
5. In 10 % of the accidents, AVs and HVs both proceeded straight prior to the collision, i.e., (Proceeding straight, Proceeding straight) (Fig. 4-(d)).
6. In 7 % of the accidents. HVs made a right turn while AVs were stopped, i.e., (Stopped, Making a right turn) (Fig. 4-(d)).
7. 70 % of the collisions occurred on HVs rear-end (Fig. 4-(e)).
8. In 13 % of the collisions, AVs and HVs sideswiped each other (Fig. 4-(e)).
9. In 50 % of the accidents, AVs and HVs were located on the same lane (83 cases, 50 %), as shown in Fig. 4-(f).
10. 92 % of the accidents took place on dry road surface (Fig. 4-(g)).
11. 90 % of the accidents occurred under abnormal road conditions (Fig. 4-(h)).
12. 77 % of the accidents happened at intersections (Fig. 4-(i)).
13. 39 % of the accidents occurred between 12 and 18 (Fig. 4-(j)).
14. 15 % of accidents occurred on red light (Fig. 4-(k)).

We had to implemented additional analytical procedures to unravel key associations between the collision types and other accident factors which could not be found otherwise with the preliminary statistical analysis.

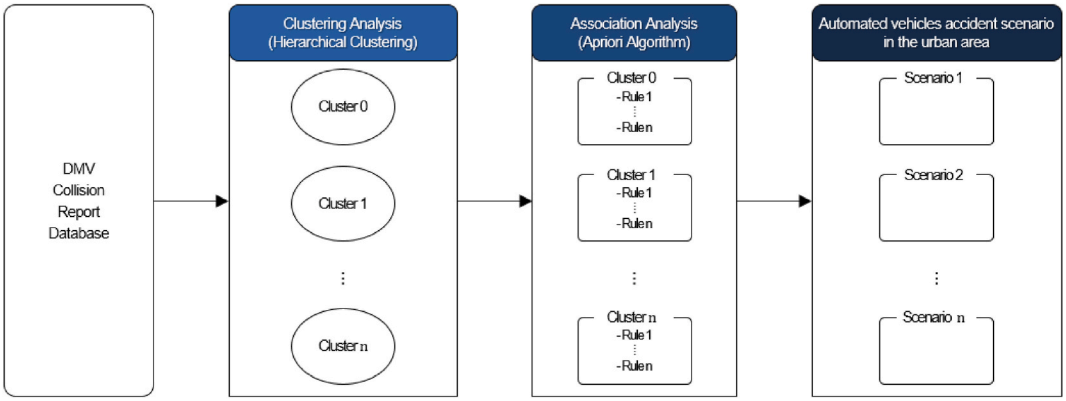


Fig. 5. Process for presenting AVs accident scenario.

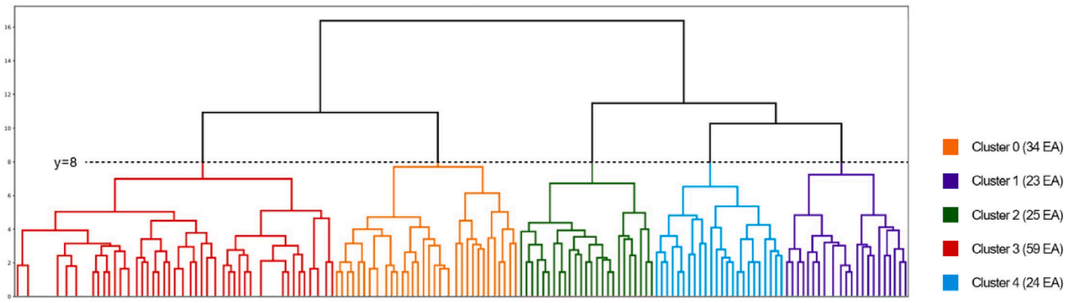


Fig. 6. Dendrogram after applying hierarchical clustering with threshold (y) set to 8.

3.4. Clustering and association rule mining

To derive a meaningful correlation between the collision types and the other accident factors, we implemented a two-phase analysis method as shown in Fig. 5.

The first phase involves a task of grouping reports exhibiting similar characteristics. To derive the groups, we ran a hierarchical clustering algorithm [47]. We chose the ward linkage method to assess the closeness between encoded reports, which computes the sum of squares of deviations between data within and across clusters [48]. By calculating the distribution of the data, we could generate clusters that are less sensitive to outliers than the non-hierarchical clustering approaches. In the second phase, we conducted association rule mining for each DMV report cluster to correlate between accident types and the other accident factors. We employed the Apriori algorithm [49] for association analysis that extracts frequent item sets according to the configurable minimum support level. With an efficient pruning of less frequent item sets, the Apriori algorithm is capable of mining meaningful association rules fast. The association rule is expressed as A(Antecedent) → B(Consequent), meaning that B co-occurred with A. In our context, A is the item set of accident factors, and B is the type of collision by AVs and HVs. How strongly A and B are correlated can be measured with the metrics defined in Eq. (1) ~ (3). Support (Eq. (1)) refers to the probability of a rule including both A and B divided by the number (N) of items in the entire set. Confidence (Eq. (2)) is the probability that B occurs when A is included in the rule, which indicates the reliability of the rule. Lift (Eq. (3)) affirms the association between A and B by computing the conditional probability of B occurring when A is in the rule divided by the probability of B occurring in the entire item sets. If there is no relation between A and B, the Lift equals 1. Otherwise, the higher the Lift value is, the more significant the relationship between A and B is.

$$\text{Support}(\text{A} \rightarrow \text{B}) = \frac{P(A \cap B)}{N} \tag{1}$$

$$\text{Confidence}(\text{A} \rightarrow \text{B}) = P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{2}$$

$$\text{Lift}(\text{A} \rightarrow \text{B}) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} \tag{3}$$

With our two-phase analysis method, association rules are mined only among the most relevant and similar collision reports within each cluster. Therefore, the association rules can be extracted significantly faster than the approach of running the Apriori algorithm

Table 2
Confirm distribution of each cluster.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Type of Accident (TA)					
Car-Alone	0 %	0 %	12 %	0 %	0 %
Car-to-Car	85 %	96 %	76 %	100 %	100 %
Movement Preceding Collision (MPC) (AVs, HVs)					
(Stopped, Proceeding straight)	38 %	13 %	8 %	53 %	13 %
Type of Collision (TC) (AVs, HVs)					
(N/A, Rear-end)	74 %	43 %	60 %	100 %	25 %
(Sideswipe, Sideswipe)	3 %	39 %	0 %	0 %	50 %
(Hit object, None)	0 %	0 %	8 %	0 %	0 %
(Head-on, Head-on)	3 %	0 %	0 %	0 %	0 %
Relative Position (RP) (AVs, HVs)					
(Left, Right)	12 %	100 %	0 %	0 %	0 %
(Right, Left)	3 %	0 %	0 %	0 %	96 %
(Same, Same)	68 %	0 %	0 %	100 %	4 %
(Opposite, Opposite)	5 %	0 %	0 %	0 %	0 %
(N/A. N/A)	12 %	0 %	100 %	0 %	0 %
Location					
Intersection	85 %	70 %	68 %	88 %	54 %
Parking lot	0 %	4 %	8 %	0 %	0 %
Road	15 %	26 %	24 %	12 %	46 %

Table 3
Common characteristics and specific accident patterns of each cluster.

Common characteristics of accident	
A rear-end collision occurred when an AV is stopped and a HV is proceeding straight due to a Car-to-Car accident at an intersection.	
Characteristics of accident for each cluster	
Cluster 0	A cluster containing an accident in which an AV and a HV faced each other and collided head-on.
Cluster 1	A cluster containing a sideswipe accident with an AV is positioned to the relatively left of a HV.
Cluster 2	A cluster containing an accident where an AV collided with an object.
Cluster 3	A cluster consisting of a rear-end accident with an AV and a HV vehicle in the same lane.
Cluster 4	A cluster containing a sideswipe accident with an AV positioned to the relatively right of a HV.

Table 4
Parameters set for association rule mining.

Index	Value			
Antecedent	TA	Weather	Lighting	AVs_MPC
	HVs_MPC	AVs_RP	HVs_RP	Location
	RS	RC	Time	TL
Consequent	AVs_TC		HVs_TC	
Min_support	0.03(3 %)			
Min_ confidence	0.7(70 %)			
Min_lift	1.5			

brute-forcefully for the entire item sets without segmenting through clustering.

4. Accident data analysis result

In this section we discuss the result of the association rule mining per collision report cluster. The rules we extracted are later sued for composing accident scenarios to test during the safety evaluation of AVs.

4.1. The result of hierarchical clustering

Our hierarchical clustering algorithm produced the five most distinct clusters when the threshold y was set to 8 in the dendrogram, as shown in Fig. 6. We computed the distribution of the appearance of every accident factor and collision type to reveal the differentiating characteristics of each cluster as shown in Table 2 and Table 3. All clusters reported an HV colliding with the rear end of a stopped HV at an intersection. Table 3 shows the unique collision patterns of each cluster. Clusters 1 and 4 were mainly about sideswipe collisions. Particularly in cluster 1, the AVs were on the left side of HVs at the moment of collision. Opposite to Cluster 1, Cluster 4 contained accidents when AVs were on the right side of HVs. Cluster 2 is the only one that contained 12 % of reports on cars

Table 5
Significant association rules extracted from each cluster (A = Antecedent, C=Consequent, S=Support, L = Lift).

ID	A	C	S	L
Cluster 1				
1	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Passing other vehicle] + [AVs_RP = Left] + [HVs_RP=Right] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[HVs_TC = Rear-end]	0.08	2.3
2	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Dark-street lights] + [AVs_MPC= Proceeding straight] + [HVs_MPC=Changing lane] + [AVs_RP = Left] + [HVs_RP= Right] + [Location = Road] + [RS = Dry] + [RC=No unusual conditions] + [Time = 18–24] + [TL = N/A]	[HVs_TC = Rear-end]	0.04	2.3
3	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Passing other vehicle] + [AVs_RP = Left] + [HVs_RP=Right] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = Red light]	[AVs_TC = Sideswipe]	0.04	2.6
4	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Passing other vehicle] + [AVs_RP = Left] + [HVs_RP=Right] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = Red light]	[HVs_TC = Sideswipe]	0.04	2.3
5	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Dark-street lights] + [AVs_MPC= Proceeding straight] + [HVs_MPC=Changing lane] + [AVs_RP = Left] + [HVs_RP= Right] + [Location = Intersection] + [RS = Dry] + [RC=N/A] + [Time = 18–24] + [TL = N/A]	[AVs_TC = Sideswipe]	0.04	2.6
6	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Dark-street lights] + [AVs_MPC= Proceeding straight] + [HVs_MPC=Changing lane] + [AVs_RP = Left] + [HVs_RP= Right] + [Location = Intersection] + [RS = Dry] + [RC=N/A] + [Time = 18–24] + [TL = N/A]	[HVs_TC = Sideswipe]	0.04	2.3
Cluster 2				
7	[TA = Car-Alone] + [Weather = Cloudy] + [Lighting = Daylight] + [AVs_MPC = Making right turn] + [HVs_MPC=N/A] + [AVs_RP=N/A] + [HVs_RP=N/A] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[AVs_TC = Hit object]	0.04	12.5
8	[TA = Car-Alone] + [Weather = Cloudy] + [Lighting = Daylight] + [AVs_MPC = Making right turn] + [HVs_MPC=N/A] + [AVs_RP=N/A] + [HVs_RP=N/A] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[HVs_TC = N/A]	0.04	6.25
9	[TA = Car-Alone] + [Weather = Cloudy] + [Lighting = Daylight] + [AVs_MPC=Changing lane] + [HVs_MPC=N/A] + [AVs_RP=N/A] + [HVs_RP=N/A] + [Location = Road] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[AVs_TC = Hit object]	0.04	12.5
10	[TA = Car-Alone] + [Weather = Cloudy] + [Lighting = Daylight] + [AVs_MPC=Changing lane] + [HVs_MPC=N/A] + [AVs_RP=N/A] + [HVs_RP=N/A] + [Location = Road] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[HVs_TC = N/A]	0.04	6.25
Cluster 3				
11	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Proceeding straight] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[AVs_TC = N/A], [HVs_TC = Rear-end]	0.08	–
12	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Proceeding straight] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = Red light]	[AVs_TC = N/A], [HVs_TC = Rear-end]	0.1	–
13	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Proceeding straight] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = Stop sign]	[AVs_TC = N/A], [HVs_TC = Rear-end]	0.03	–
14	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Proceeding straight] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = N/A]	[AVs_TC = N/A], [HVs_TC = Rear-end]	0.07	–
15	[TA = Car-to-Car] + [Weather = Cloudy] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC=Proceeding straight] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = N/A]	[AVs_TC = N/A], [HVs_TC = Rear-end]	0.03	–
16	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Stopped] + [HVs_MPC = Making right turn] + [AVs_RP=Same] + [HVs_RP=Same] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = Green light]	[AVs_TC = N/A], [HVs_TC	0.03	–

(continued on next page)

Table 5 (*continued*)

ID	A	C	S	L
		= Rear-end]		
Cluster 4				
17	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Proceeding straight] + [HVs_MPC=Other unsafe turning] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = N/A]	[HVs_TC = Head-on]	0.04	8
18	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Proceeding straight] + [HVs_MPC=Changing lane] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[AVs_TC = N/A]	0.04	3
19	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Proceeding straight] + [HVs_MPC=Changing lane] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[HVs_TC = Rear-end]	0.04	4
20	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Dark-street lights] + [AVs_MPC = Making right turn] + [HVs_MPC = Making right turn] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 18–24] + [TL = N/A]	[AVs_TC = Sideswipe]	0.04	1.85
21	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Dark-street lights] + [AVs_MPC = Making right turn] + [HVs_MPC = Making right turn] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 18–24] + [TL = N/A]	[HVs_TC = Sideswipe]	0.04	2
22	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Proceeding straight] + [HVs_MPC=Passing other vehicle] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[AVs_TC = Sideswipe]	0.04	1.85
23	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [AVs_MPC=Proceeding straight] + [HVs_MPC=Passing other vehicle] + [AVs_RP=Right] + [HVs_RP = Left] + [Location = Intersection] + [RS = Dry] + [RC=No unusual conditions] + [Time = 6–12] + [TL = N/A]	[HVs_TC = Sideswipe]	0.04	2
24	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [Avs_MPC=Changing lane] + [HVs_MPC=Changing lane] + [Avs_RP=Right] + [HVs_RP = Left] + [Location = Road] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = N/A]	[Avs_TC = Sideswipe]	0.04	1.85
25	[TA = Car-to-Car] + [Weather = Clean] + [Lighting = Daylight] + [Avs_MPC=Changing lane] + [HVs_MPC=Changing lane] + [Avs_RP=Right] + [HVs_RP = Left] + [Location = Road] + [RS = Dry] + [RC=No unusual conditions] + [Time = 12–18] + [TL = N/A]	[HVs_TC = Sideswipe]	0.04	2

colliding with non-vehicular objects by themselves. Cluster 3 only had HV hitting AVs' read-end on the same lane. Cluster 0 is the only one that contained head-on collisions between AVs and HVs that were moving in the opposite direction.

4.2. The result of association analysis

We set minimum support, minimum confidence, and minimum lift to 0.03, 0.7, and 1.5, respectively. Antecedents and consequents are specified in Table 4.

As a result, we discovered a total of 313,748 association rules. We compiled a list of the most significant rules in terms of support and lift. These are the rules in which all accident factors are included in the antecedent, as shown in Table 5. In cluster 0, our algorithm did not derive any significant association rules. Cluster 1, on the other hand revealed six association rules, such as the one with rule #3. The rule states that an HV sideswiped a stopped AV on the left while the HV passed other vehicles at the intersection on the red light during the day under clean weather.

Cluster 2 mostly shows AVs alone hitting non-vehicular objects while changing lanes or making right turns (rule #7 and rule #9).

From Cluster 3, all the accidents were rear-ended collisions. Thus, the lift could not be computed. Instead, we extracted rules with a support value of 0.03 or higher. For instance, rule #11 indicates that HVs frequently hit stopped AVs on the same lane and the intersections with a dry surface. Considering rule #14, we can infer that the collision pattern expressed in rule #11 frequently occurred during the daytime.

Cluster 4 mostly shows the situations leading to rear-end or sideswipe collisions. In most cases in this cluster, AVs were on the right lane proceeding straight while HVs on the left-hand side changed lanes.

We could not find frequent accident patterns regarding road surface and road conditions. Reports about accidents at intersections specify the existence of traffic lights. However, the detailed state of the traffic lights was omitted. Note that the rules extracted by the association analysis unravel a more detailed correlation between the accident factors and accident types.

5. Deriving AV accident scenarios on urban areas

5.1. Scenario composition framework and accident factor selection

From the frequently seen accident patterns revealed through our data analysis method, we generate accident scenarios for AV safety tests in urban areas. As mentioned earlier, several projects, such as PEGASUS, CETRAN, and ENABLE-S [39,50], have introduced various scenarios to evaluate the safety of AVs. The PEGASUS project systematizes methods and requirements for securing the safety of automated driving functions and specifies accident scenarios that may occur on highways. In addition, PEGASUS project suggests scenarios at different levels of abstraction such as functional, logical, and concrete scenarios based on a 6-layer model [16]. The 6-layer

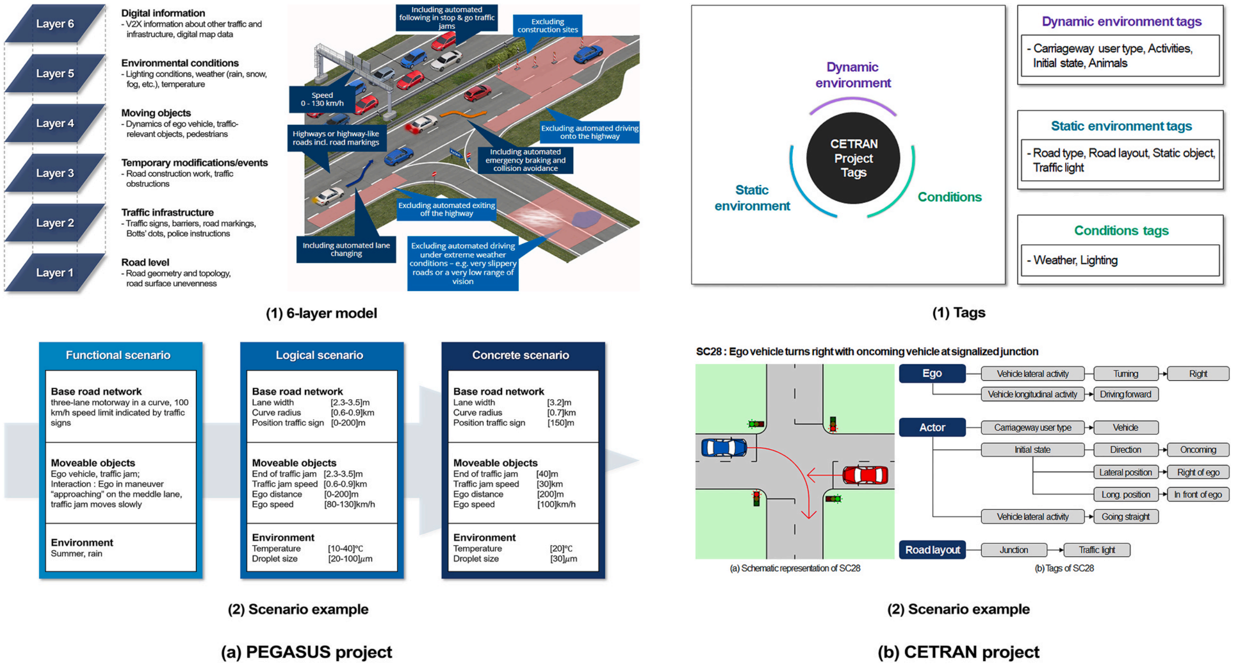


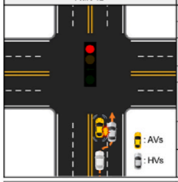
Fig. 7. Information described in the accident scenarios composed by PEGASUS and CETRAN projects.

Table 6
Accident factors extracted from DMV collision report and their appearance in PEGASUS or CETRAN projects. 11 factors in bold face are the factors that appear in PEGASUS or CETRAN project. We use all 14 factors for composing scenarios.

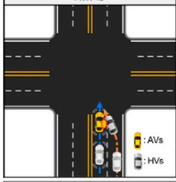
Accident Factors used in the DMV Reports	Appearance of the Accident Factors								
	PEGASUS Project Layers						CETRAN Project Tags		
	1	2	3	4	5	6	Dynamic Environment	Static Environment	Conditions
Type of Accident							O		
Weather					O				O
Lighting					O				O
AVs Movement Preceding Collision				O			O		
HVs Movement Preceding Collision				O			O		
AVs Type of Collision									
HVs Type of Collision									
AVs Relative Position				O			O		
HVs Relative Position				O			O		
Location	O							O	
Road Surface	O								
Road Condition			O						
Time									
Traffic Light	O							O	

model categorizes six *ensembles*: road level, traffic infrastructure, temporary modifications/events, moving objects, environmental conditions, and digital information. The ensemble is subdivided into *factors*, including road geometry, traffic signals, road conditions, dynamics of the ego vehicle, lighting conditions, and V2X information. These factors are essential for highway scenarios, forming specific highway scenarios for evaluating the safety of AV. A functional scenario is a text description of road networks, stationary or non-stationary objects, and environmental conditions. The functional scenario also provides lane width, speed limit, vehicle movement, and weather information, as shown in Fig. 7-(a). Based on the factors defined by the functional scenario, the logical scenario sets the range of values, and the concrete scenario sets the value of individual factors.

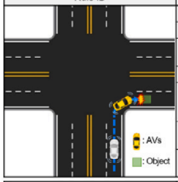
The CETRAN project [51] specifies 64 representative scenarios to evaluate AVs' safety based on the NHTSA pre-crash scenarios [52]. CETRAN presents moving objects as a dynamic environment tag and stationary objects as a static environment tag, as shown in Fig. 7-(b). Also, CETRAN specifies condition tags such as weather and lighting. Fig. 7-(b) shows an example of the scenario provided by CETRAN. The scenario contains a schematic diagram to identify the situation. Also, it provides detailed information on ego vehicle, actor vehicle, and road layout, based on the three tags (dynamic environment, state environment, and conditions) to evaluate the safety of AV.



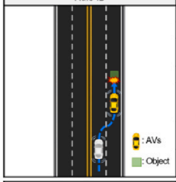
Name	Scenario 1		
Cluster information	Cluster 1		
Rule ID	3 and 4		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	Red light
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Passing other vehicle
AVs Relative Position	Left	HVs Relative Position	Right
AVs Type of Collison	Sideswipe	HVs Type of Collison	Sideswipe



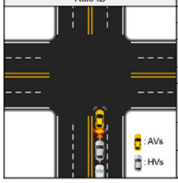
Name	Scenario 2		
Cluster information	Cluster 1		
Rule ID	5 and 6		
TA	Car-to-Car	Weather	Clean
Lighting	Dark-street lights	Location	Intersection
Road Surface	Dry	Road Condition	N/A
Time	18~24	Traffic Light	N/A
AVs Movement Preceding Collision	Proceeding straight	HVs Movement Preceding Collision	Changing lane
AVs Relative Position	Left	HVs Relative Position	Right
AVs Type of Collison	Sideswipe	HVs Type of Collison	Sideswipe



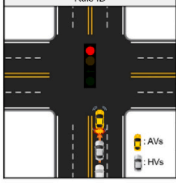
Name	Scenario 3		
Cluster information	Cluster 2		
Rule ID	7 and 8		
TA	Car-Alone	Weather	Cloudy
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	N/A
AVs Movement Preceding Collision	Making right turn	HVs Movement Preceding Collision	N/A
AVs Relative Position	N/A	HVs Relative Position	N/A
AVs Type of Collison	Hit object	HVs Type of Collison	N/A



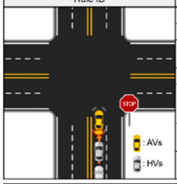
Name	Scenario 4		
Cluster information	Cluster 2		
Rule ID	9 and 10		
TA	Car-Alone	Weather	Clean
Lighting	Daylight	Location	Road
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	N/A
AVs Movement Preceding Collision	Changing lane	HVs Movement Preceding Collision	N/A
AVs Relative Position	N/A	HVs Relative Position	N/A
AVs Type of Collison	Hit object	HVs Type of Collison	N/A



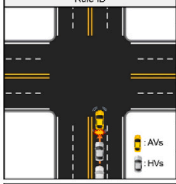
Name	Scenario 5		
Cluster information	Cluster 3		
Rule ID	11		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	N/A
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Proceeding straight
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



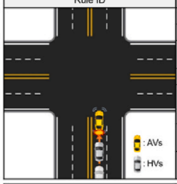
Name	Scenario 6		
Cluster information	Cluster 3		
Rule ID	12		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	Red light
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Proceeding straight
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



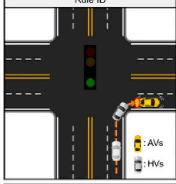
Name	Scenario 7		
Cluster information	Cluster 3		
Rule ID	13		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	Stop sign
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Proceeding straight
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



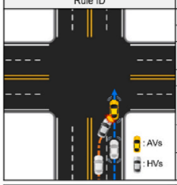
Name	Scenario 8		
Cluster information	Cluster 3		
Rule ID	14		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	N/A
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Proceeding straight
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



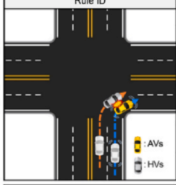
Name	Scenario 9		
Cluster information	Cluster 3		
Rule ID	15		
TA	Car-to-Car	Weather	Cloudy
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	N/A
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Proceeding straight
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



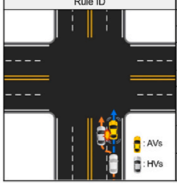
Name	Scenario 10		
Cluster information	Cluster 3		
Rule ID	16		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	Green light
AVs Movement Preceding Collision	Stopped	HVs Movement Preceding Collision	Making right turn
AVs Relative Position	Same	HVs Relative Position	Same
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



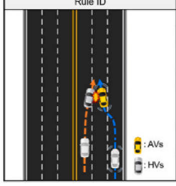
Name	Scenario 11		
Cluster information	Cluster 4		
Rule ID	18 and 19		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	N/A
AVs Movement Preceding Collision	Proceeding straight	HVs Movement Preceding Collision	Changing lane
AVs Relative Position	Right	HVs Relative Position	Left
AVs Type of Collison	N/A	HVs Type of Collison	Rear-end



Name	Scenario 12		
Cluster information	Cluster 4		
Rule ID	20 and 21		
TA	Car-to-Car	Weather	Clean
Lighting	Dark-street lights	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	18~24	Traffic Light	N/A
AVs Movement Preceding Collision	Making right turn	HVs Movement Preceding Collision	Making right turn
AVs Relative Position	Right	HVs Relative Position	Left
AVs Type of Collison	Sideswipe	HVs Type of Collison	Sideswipe



Name	Scenario 13		
Cluster information	Cluster 4		
Rule ID	22 and 23		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Intersection
Road Surface	Dry	Road Condition	No unusual condition
Time	6~12	Traffic Light	N/A
AVs Movement Preceding Collision	Proceeding straight	HVs Movement Preceding Collision	Passing other vehicle
AVs Relative Position	Right	HVs Relative Position	Left
AVs Type of Collison	Sideswipe	HVs Type of Collison	Sideswipe



Name	Scenario 14		
Cluster information	Cluster 4		
Rule ID	24 and 25		
TA	Car-to-Car	Weather	Clean
Lighting	Daylight	Location	Road
Road Surface	Dry	Road Condition	No unusual condition
Time	12~18	Traffic Light	N/A
AVs Movement Preceding Collision	Changing lane	HVs Movement Preceding Collision	Changing lane
AVs Relative Position	Right	HVs Relative Position	Left
AVs Type of Collison	Sideswipe	HVs Type of Collison	Sideswipe

Fig. 8. Accident scenarios generated based on the association rules.

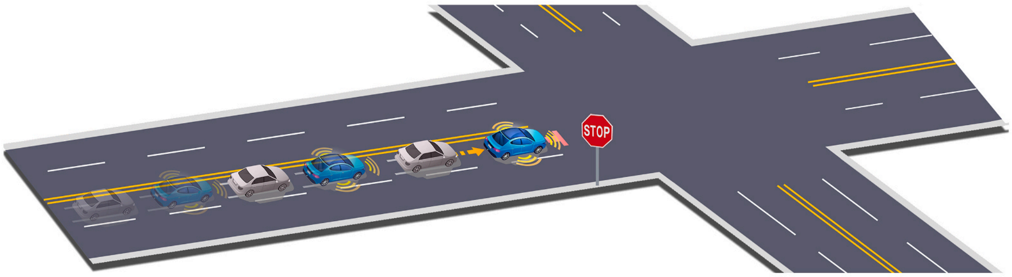
Safety Function Requirement	Recognition and response evaluation for the Target vehicle approaching from behind				
Type of Accident	Car-to-Car				
Type of Collision	Rear-end				
Weather	Clean				
Lighting	Daylight				
Time	12~18				
Location	Intersection				
Visualization of the Accident					
Text Description of Accident	In a situation where the ego vehicle(AVs) is stopping at an intersection, the ego vehicle(AVs) should respond to the situation where the target vehicle(HVs) approaches without securing a safe distance.				
Road Geometry	Number of lanes	2			
	Road alignment	Straight line			
	Road surface	Dry			
	Road condition	No unusual condition			
	Traffic light	Stop sign			
Object Movement	Object	Vehicle Type	Driving Lane	Relative position	Movement
	Ego Vehicle (AVs)	Passenger Car	1st lane	-	Stopped
	Target Vehicle (HVs)	Passenger Car	1st lane	Behind	Proceeding straight
Source of Scenario	Data Source	DMV collision report in California			

Fig. 9. Scenario #7 written in a PEGASUS functional scenario format.

In Table 6, we listed 14 accident factors that can be extracted through the DMV collision report. We found out that both PEGASUS and CETRAN scenarios specify the accident factors such as the movements of AVs and HVs preceding collision, traffic light information, and the relative positions of the vehicles. Unlike CETRAN, the PEGASUS scenarios specify road surface and condition. However, PEGASUS scenarios do not specify the type of accidents, while CETRAN scenarios do.

Our accident scenarios describe AV accidents more thoroughly by including the accident factors that are considered neither by PEGASUS nor by CETRAN.

The AVs urban accident scenario was presented using the association rule, including all 14 accident factors. Given the association rules with statistical significance (min_support = 0.03 or min_lift = 1.5), we combine the association rules to constitute scenarios. Any two association rules can be combined if they satisfy three conditions as follows:

(1) Condition 1: An antecedent must contain all 14 accident factors.
(2) Condition 2: The two association rules can be combined if they have identical antecedents.
(3) Condition 3: The pair of consequents from the association rules generated by Rule 2 must be physically plausible and match the accident situation described in the antecedents. For example, if the accident was not a car-to-car collision, then having an AV colliding with a stationary object and an HV sideswiping the AV is not physically plausible. As another example, it is not physically plausible to state that HV was in a head-on collision with the AV while AV was reported to have sideswiped the HV.

Except for rules #1, #2, and #17, an association rule in Table 5 matched another rule to constitute one of the 14 accident scenarios we have compiled in Fig. 8. In each scenario, we have provided a visual depiction of the vehicle motions leading to collisions. Besides the illustrating the accidents, we have provided the 14 factors explaining the accident situation and the collision types. We have

SECTION 5 – ACCIDENT DETAILS - DESCRIPTION

■ Autonomous Mode ☐ Conventional Mode

On December 77, 2021 at 10:47 AM PST a Waymo Autonomous Vehicle("Waymo AV") operating in San Francisco, CA was in a collision involving an SUV at Oak Street at Stanyan Street.
While in autonomous mode, the Waymo AV came to a stop at a stop sign on Oak Street. While the ADV was stopped, waiting for a pedestrian to cross and for traffic on Stanytan Street to clear, an SUV approached the Waymo AV from behind and made contact with the rear of the Waymo AV. At the time of the impact, the Waymo AV's Level 4 ADS was engaged in autonomous mode, and a test driver was present (in the driver's seating position). The Waymo AV sustained minor damage to the rear bumper.

(a) DMV Collision Report (Original Version)

SECTION 5 – ACCIDENT DETAILS - DESCRIPTION

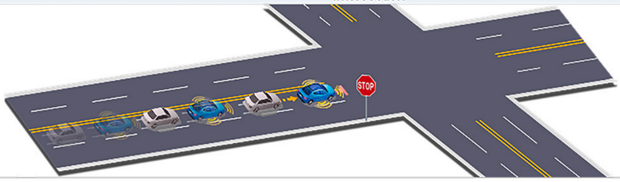
1. Essential elements (Check if the elements are described below)

☐ Mode of AVs	☐ Longitudinal movement of AVs	☐ Longitudinal movement of HVs
☐ Lateral movement of AVs	☐ Lateral movement of HVs	☐ Longitudinal speed of AVs
☐ Longitudinal speed of HVs	☐ Driving lane of AVs	☐ Driving lane of HVs
☐ Longitudinal position of AVs	☐ Longitudinal position of HVs	☐ Numbers of target vehicles

2. Detailed description of the situation

In this situation ~

(b) DMV Collision Report (Suggested Revision)

Safety Function Requirement	Recognition and response evaluation for the Target vehicle approaching from behind						
Type of Accident	Car-to-Car						
Type of Collision	Rear-end						
Weather	Clean						
Lighting	Daylight						
Time	12~18						
Location	Intersection						
Visualization of the Accident Situation							
Text Description of Accident	In a situation where the ego vehicle(AVs) is stopping at an intersection, the ego vehicle(AVs) should respond to the situation where the target vehicle(HVs) approaches without securing a safe distance.						

Layer	Item	Element	Description	Variable values: Data type	Min. value	Max. value	Δ
Layer 1 : Plane data	Road section	Road section type	-	Categorical	Road, Shoulder, Flank, Deceleration lane, Acceleration lane		
	Road alignment	Road alignment type	-	Categorical	Straight, Curve, Hair pin curve, Transition curve or section		
		Minimum plane curve radius	Automatically determined according to the road design speed and longitudinal slope.	Integer	-	-	-
		Minimum plane curve length	Automatically determined according to the road intersection angle and road design speed.	Integer	-	-	-
	Road slope	Maximum longitudinal slope	-	Float	-6%	6%	1%
	Road surface	Road surface type	-	Categorical	Dry, Wet, Snowy-Icy, Slippery(Muddy, Oily, etc.)		
	Road condition	Road condition type	-	Categorical	Holes, Loose material, Obstruction, Construction-repair zone, Reduced roadway width, Flooded, No unusual conditions		
Layer 4 : Scenario participant data	Ego Vehicle	Vehicle type	-	Categorical	Passenger car, Van, Bus, Truck, Emergency car, Motorcycle, Others		
		Initial driving lane	-	Categorical	1	6	1
		Initial movement	Lateral movement	Integer	Proceeding straight, Cut-in, Cut-out, Cut-through, Others		
		Driving lane in the case of situation	Longitudinal movement	Integer	Constant speed, Accelerating, Decelerating, Stopping, Others		
		Movement in the case of situation	Lateral movement	Categorical	Proceeding straight, Cut-in, Cut-out, Cut-through, Others		
			Longitudinal movement	Categorical	Constant speed, Accelerating, Decelerating, Stopping, Others		
		Longitudinal speed in the case of situation	The maximum value depends on the real road speed limit	Integer	10km/h	30~60km/h	10
	Target Vehicle	Number of vehicle	2	Integer	0	5	1
		Vehicle type	-	Categorical	Passenger car, Van, Bus, Truck, Emergency car, Motorcycle, Others		
		Whether vehicle is AV	None	Categorical	Level 0, Level 1, Level 2, Level 3, Level 4, Level 5		
		Initial driving or located lane	-	Integer	1	6	1
		Initial relative location for the Ego	-	Categorical	ahead, ahead-left, ahead-right, side-left, side-right, behind, behind-left, behind-right, oncoming, oncoming-left, oncoming-right		
		Initial movement	Lateral movement	Categorical	Proceeding straight, Cut-in, Cut-out, Cut-through, Others		
			Longitudinal movement	Categorical	Constant speed, Accelerating, Decelerating, Stopping, Others		

(c) Logical & Concrete scenario

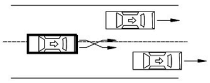
Fig. 10. An examples of current DMV report and suggested revision. A logical and concrete scenario for Scenario #7 based on PEGASUS scenario description framework.

indicated the rules that were combined to constitute a given scenario. We have also specified the cluster to which the combined rules belong. The scenarios generated from Cluster 3 are associated with a single association rule because all the rules describe the situation where HVs hit the rear end of the AVs.

We presented Scenario #7 from Fig. 8 as a functional scenario set by the PEGASUS project, as shown in Fig. 9. The safety function requirement section is about the evaluation element of AVs. Scenario #7 is to test the response of the stopped ego vehicle when the

Following Vehicle Making a Maneuver and Approaching Lead Vehicle

Typical Scenario: Vehicle is changing lanes or passing in an urban area, in daylight, under clear weather conditions, at a non-junction with a posted speed limit of 55 mph; and closes in on a lead vehicle.



Factor Over-Representation: Intersection-related location, inattention, speeding, and younger driver are over-represented (based on a simple comparison of percentages).

Dynamic Variations: Vehicle is turning right and then closes in on a lead vehicle (22% of crashes).

Scenario Severity: Table below quantifies the annual severity of this crash scenario in terms of five different metrics based on 2004 GES statistics. This table also provides the ratios of people involved by maximum injury severity using the KABCO and AIS injury scales. About 0.50 percent of all people involved in this crash scenario suffered high-level MAIS 3+ injuries (serious, severe, critical, or fatal).

Crash Severity		Scenario	Scenario/All
No. of crashes		85,000	1.44%
No. of vehicles involved		180,000	1.69%
No. of people involved		249,000	1.66%
Societal Cost	Economic cost	$1,212,000,000	1.01%
	Functional years lost	18,000	0.67%
KABCO Injury Scale	None	0.860	1.052
	Possible	0.103	0.946
	Non-incapacitating	0.023	0.482
	Incapacitating	0.009	0.487
	Fatal	0.0001	0.053
	Unknown	0.004	1.049
	Died prior	-	-
AIS Injury Scale	None	0.817	1.047
	Minor	0.163	0.864
	Moderate	0.015	0.707
	Serious	0.004	0.632
	Severe	0.0005	0.573
	Critical	0.0002	0.516
	Fatal	0.0001	0.053
	Injured people per crash	0.533	0.962

(a) Typical scenario of NHTSA

Following Vehicle Making a Maneuver and Approaching Lead Vehicle

Driving Environment

Lighting	Daylight	76%
	Dark Lighted	16%
	Dark	3%
	Dawn/Dusk	4%
Weather	Clear	91%
	Adverse	9%
Road Surface	Dry	85%
	Wet/Slippery	15%
Road Alignment	Straight	84%
	Curve	16%
Road Profile	Level	80%
	Other	20%
Land Use	Rural	42%
	Urban	58%
Day	Weekday	77%
	Weekend	23%
Relation to Roadway	On Roadway	96%
	Shoulder/Parking Lane	1%
	Off Roadway	2%
	Left Turn Lane	0.3%
	Unknown	-
Relation to Junction	Non-Junction	36%
	Intersection	7%
	Intersection-Related	33%
	Driveway/Alley	4%
	Entrance/Exit Ramp	6%
	Rail Grade Crossing	0.3%
	Other/Unknown	14%
Posted Speed Limit (mph)	<= 20	0.5%
	25	8%
	30	9%
	35	25%
	40	10%
	45	19%
	50	5%
	>= 55	24%
Traffic Control Device	No Traffic Controls	50%
	Traffic Signal	29%
	Stop/Yield Sign	14%
	Other	7%

Driver

Alcohol	Yes	5%
	No	95%
Vision Obscured	No Obstruction	64%
	Vision Obscured	2%
	Unknown	34%
Driver Distracted	Inattention	29%
	Sleepy	0.3%
	Not Distracted	24%
	Unknown	47%
Speeding	Yes	25%
	No	64%
	Unknown	11%
Violation	Speeding	-
	Reckless	1%
	None	44%
	Other	42%
	Unknown	13%
Impairment	Ill/Blackout	0.1%
	Drowsy	0.2%
	None	88%
	Other	2%
	Unknown	10%
Gender	Male	59%
	Female	41%
Age	Younger <= 24	33%
	Middle = 25 to 64	62%
	Older >= 65	5%

Vehicle

Contributing Factors	Yes	1%
	No	80%
	Unknown	20%
Rollover	Yes	0.1%
	No	100%
Pre-Event Movement	No Driver Present	-
	Going Straight	-
	Decelerating in Traffic Lane	-
	Accelerating in Traffic Lane	-
	Starting in Traffic Lane	-
	Stopped in Traffic Lane	-
	Passing Another Vehicle	9%
	Parked in Travel Lane	-
	Leaving a Parked Position	6%
	Entering a Parked Position	1%
	Turning Right	22%
	Turning Left	7%
	Making U-turn	0.3%
	Backing Up	-
	Negotiating a Curve	-
	Changing Lanes	36%
	Merging	4%
	Prior Corrective Action	3%
	Other	12%
Driver Avoidance Maneuver	Object in Road	0.2%
	Poor Road Conditions	-
	Animal in Road	-
	Vehicle in Road	12%
	Non-Motorist in Road	-
	Hit and Run	17%
	No Driver Present	-
	Other Avoidance Maneuver	-
	Unknown	57%
	None	13%
	Phantom Vehicle	0.01%
Corrective Action Attempted	No Driver Present	-
	No Avoidance Maneuver	11%
	Braking	5%
	Releasing Brakes	-
	Steering	8%
	Braked and Steered	1%
	Accelerated	0.2%
	Accelerated and Steered	-
	Other	0.1%
	Unknown	75%

Driver and vehicle statistics represent the striking light vehicle.

(b) Probabilities of dynamic factors

Fig. 11. NHTSA pre-crash scenario.

target vehicle approaches from behind. This scenario should be tested at the intersection in clean weather and daylight. In addition, the scenario contains a visual and textual description of the accident. Road geometry is a section that provides information about the number of lanes, road alignment, road surface, road condition, and traffic lights. The object movement section provides information on the vehicle types, driving positions, and movements of vehicles during critical situations. The source of the scenario is a section presenting the source of data we utilized to generate the scenario. The example in Fig. 9 states that the scenario was created using the DMV collision report in California.

The logical and concrete scenario elaborates more on the functional scenario, as shown in Fig. 10-(c). The framework is proposed by Ko et al. (2022) [53] and Kang et al. (2022a) [39] based on the PEGASUS project. The logical scenario provides a range of factors. In contrast, the concrete scenario specifies the exact values of the factors to configure for the safety evaluation of AVs. Fig. 10-(a) shows the sample DMV report that tends to be inconsistently formatted by the car manufacturers. Fig. 10-(b) shows the suggested revision of the DMV report format that conforms to the logical and concrete scenarios. We have added essential description elements such as AVs mode, vehicle speed, vehicles' longitude movement, the number of lanes, and the number of the target vehicle. These elements can be fully described in Section 5.2 of the DMV report.

5.2. *Comparison between AV and HV accident scenarios*

To delve into the peculiar AV accident patterns not normally seen by conventional HVs, we compared the AV accident scenarios described in section 5.1 with 37 HV-only pre-crash scenarios generated by NHTSA. The NHTSA scenarios are based on the 2004 General Estimates System crash database. It provides specific information on the typical scenario (Fig. 11-(a)) and provides the probability of the influential factors in the scenario (Fig. 11-(b)).

It is necessary to AV and HV accident patterns under the same conditions. NHTSA and CA DMV data both consist of traffic accidents caused by HVs. We calculated the probability of HVs getting involved in an accident under the NHTSA scenarios similar to ours.

We selected four typical scenarios that are similar to our scenarios in NHTSA, Following Vehicle Making a Maneuver and Approaching Lead Vehicle (scenario #1, #2, #11, #12, and #13), Vehicle Contacting Object Without Prior Vehicle Maneuver (scenario #3 and #4), Following Vehicle Approaching a Stopped Lead Vehicle (Scenario #5 to #10) and Vehicle Changing Lanes-Vehicle Traveling in Same Direction (scenario #14). In Table 7, HV/HV represents the factors and probabilities of conventional HVs accidents as presented in the existing NHTSA. On the other hand, HV/AV represents the factors and probabilities of involving HVs and AVs. The probability of HV/HV and HV/AV in scenario #1 has 2.455 % and 0.356 %, respectively. The two cases differed in terms of Traffic Control Device and Pre-Event Movement. Likewise, if they have the differences in the probability between HV/HV and HV/AV, we

Table 7
Probability of our AV scenarios occurring in NHTSA HV pre-crash scenario.

Scenario	Typical scenario typology (NHTSA)	Type	Lighting	Weather	Road Surface	Land Use	Relation to Junction	Traffic Control Device	Pre-Event Movement	Probability
1	Following Vehicle Making a Maneuver and Approaching Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	*No traffic controls*	*Changing lanes*	2.455 %
			0.76	0.91	0.85	0.58	0.4	*0.5*	*0.36*	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	*Traffic signal*	*Passing another vehicle*	0.356 %
			0.76	0.91	0.85	0.58	0.4	*0.29*	*0.09*	
2	Following Vehicle Making a Maneuver and Approaching Lead Vehicle	HV/ HV	*Daylight*	Clear	Dry	Urban	Intersection	No traffic controls	Changing lanes	2.455 %
			0.76	0.91	0.85	0.58	0.4	0.5	0.36	
		HV/ AV	*Dark lighted*	Clear	Dry	Urban	Intersection	No traffic controls	Changing lanes	0.517 %
			0.16	0.91	0.85	0.58	0.4	0.5	0.36	
3	Vehicle Contacting Object Without Prior Vehicle Maneuver	HV/ HV	Daylight	*Clear*	Dry	Urban	*Non-junction*	No traffic controls	*Other*	4.463 %
			0.46	*0.87*	0.64	0.66	*0.7*	0.82	*0.46*	
		HV/ AV	Daylight	*Adverse*	Dry	Urban	*Intersection*	No traffic controls	*Turning right*	0.026 %
			0.46	*0.13*	0.64	0.66	*0.14*	0.82	*0.09*	
4	Vehicle Contacting Object Without Prior Vehicle Maneuver	HV/ HV	Daylight	Clear	Dry	Urban	Non-junction	No traffic controls	*Other*	4.463 %
			0.46	0.87	0.64	0.66	0.7	0.82	*0.46*	
		HV/ AV	Daylight	Clear	Dry	Urban	Non-junction	No traffic controls	*Changing lanes*	0.097 %
			0.46	0.87	0.64	0.66	0.7	0.82	*0.01*	
5	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	0.45	0.77	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	0.45	0.77	
6	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	*No traffic controls*	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	*0.45*	0.77	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	*Traffic signal*	Going straight	4.498 %
			0.81	0.85	0.79	0.51	0.54	*0.39*	0.77	
7	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	*No traffic controls*	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	*0.45*	0.77	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	*Stop/yield sign*	Going straight	1.038 %
			0.81	0.85	0.79	0.51	0.54	*0.09*	0.77	
8	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	0.45	0.77	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Going straight	5.19 %
			0.81	0.85	0.79	0.51	0.54	0.45	0.77	
9	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	*Clear*	Dry	Urban	Intersection	No traffic controls	Going straight	5.19 %
			0.81	*0.85*	0.79	0.51	0.54	0.45	0.77	
		HV/ AV	Daylight	*Adverse*	Dry	Urban	Intersection	No traffic controls	Going straight	0.916 %
			0.81	*0.15*	0.79	0.51	0.54	0.45	0.77	
10	Following Vehicle Approaching a Stopped Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	*No traffic controls*	*Going straight*	5.19 %
			0.81	0.85	0.79	0.51	0.54	*0.45*	*0.77*	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	*Traffic signal*	*Turning right*	0.003 %
			0.81	0.85	0.79	0.51	0.54	*0.39*	*0.0005*	
11	Following Vehicle Making a Maneuver and Approaching Lead Vehicle	HV/ HV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Changing lanes	2.455 %
			0.76	0.91	0.85	0.58	0.4	0.5	0.36	
		HV/ AV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	Changing lanes	2.455 %

(*continued on next page*)

Table 7 (*continued*)

Scenario	Typical scenario typology (NHTSA)	Type	Lighting	Weather	Road Surface	Land Use	Relation to Junction	Traffic Control Device	Pre-Event Movement	Probability
			0.76	0.91	0.85	0.58	0.4	0.5	0.36	
12	Following Vehicle Making a Maneuver and Approaching Lead Vehicle	HV/HV	*Daylight*	Clear	Dry	Urban	Intersection	No traffic controls	*Changing lanes*	2.455 %
			0.76	0.91	0.85	0.58	0.4	0.5	*0.36*	
		HV/AV	*Dark lighted*	Clear	Dry	Urban	Intersection	No traffic controls	*Turning right*	0.316 %
			0.16	0.91	0.85	0.58	0.4	0.5	*0.22*	
13	Following Vehicle Making a Maneuver and Approaching Lead Vehicle	HV/HV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	*Changing lanes*	2.455 %
			0.76	0.91	0.85	0.58	0.4	0.5	*0.36*	
		HV/AV	Daylight	Clear	Dry	Urban	Intersection	No traffic controls	*Passing another vehicle*	0.614 %
			0.76	0.91	0.85	0.58	0.4	0.5	*0.09*	
14	Vehicle Changing Lanes-Vehicle Traveling in Same Direction	HV/HV	Daylight	Clear	Dry	Urban	Non-junction	No traffic controls	Changing lanes	11.102 %
			0.74	0.89	0.83	0.54	0.69	0.79	0.69	
		HV/AV	Daylight	Clear	Dry	Urban	Non-junction	No traffic controls	Changing lanes	11.102 %
			0.74	0.89	0.83	0.54	0.69	0.79	0.69	

highlighted them in Table 7. For example, in scenario #2, HV/AV occurred during Dark-lighted while HV/HV tends to occur more during daylight. The probability of HV/AV in scenario #3, #4, and #10 was significantly low, below 0.1 %, whereas the probabilities of HV/HV were orders of magnitude higher in those scenarios.

These findings highlight two important points:

While some AV-derived scenarios may appear similar to typical scenarios involving conventional HVs, such as scenario #5, #8, #11, and #14, most AV scenarios exhibit significant differences. These probability gaps represent edge cases that may have been overlooked from the perspective of HVs. To prevent AV traffic accidents resulting from these edge cases, it is crucial to conduct safety assessments based on our study.

The utilization of AV data is vital for generating accurate AV scenarios, as the accident patterns of HVs and AVs differ. This discovery is invaluable, emphasizing the need to test AVs under different accident scenarios than HVs. Conducting a more in-depth analysis to understand the reasons behind the distinctive accident patterns exhibited by AVs would be an interesting avenue for future research.

It is important to note that the NHTSA did not consider significant factors that can influence traffic accidents, such as relative vehicle positions.

From the perspective of AVs, acquiring accurate information on the surrounding vehicle positions is a fundamental requirement for facilitating Advanced Driving Assistance Systems. Human drivers can perceive surrounding vehicles. However, their level of awareness is less extensive than that of AVs. Therefore, to comprehensively assess factors influencing accidents involving AVs, it is essential to study AV accident data.

6. Conclusion

We used clustering and association rule mining techniques to group similar and statistically significant patterns of AV accidents in urban areas with more challenging driving conditions than highways for the AVs.

We collected raw data from 370 accidents reported by the DMV of California, USA. We obtained six different clusters of accidents and 313,748 association rules. With minimum support or lift of 0.03 and 1.5, respectively, we could narrow down to 25 association rules that can constitute an accident scenario for AV safety tests. We have provided a novel method for combining any two association rules to derive functional, logical, and concrete scenarios. We have extended the PEGASUS scenario description framework to include detailed collision types. In addition, we suggested revising the current DMV report to contain a complete description of the accident situation.

We have derived 14 scenarios significantly different from the conventional HV accident scenarios reported by NHTSA. Such a discovery urges AVs to be reliably tested under more relevant scenarios than those involving only HVs.

Data availability statement

The authors do not have permission to share the data used in this research.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Hojun Lee: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Minhee Kang:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - review & editing. **Keeyeon Hwang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing. **Young Yoon:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 22AMDP-C161754-02)

References

[1] NSTC, U., Ensuring American leadership in AVs technologies: AVs 4.0. Las Vegas, Recuperado el 25 (2020), 2020-02.
[2] SAE, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles Automated Driving Systems, 2014, p. J3016.
[3] H. Lee, K. Hwang, M. Kang, J. Song, Black ice detection using CNN for the Prevention of Accidents in AVs, in: 2020 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2020, December, pp. 1189–1192, https://doi.org/10.1109/CSCI51800.2020.00222.
[4] M. Kang, J. Song, K. Hwang, For preventative automated driving system (PADS): traffic accident context analysis based on deep neural networks, Electronics 9 (11) (2020) 1829, https://doi.org/10.3390/electronics9111829.
[5] Q. Liu, X. Wang, X. Wu, Y. Glaser, L. He, Crash comparison of autonomous and conventional vehicles using pre-crash scenario typology, Accid. Anal. Prev. 159 (2021) 106281, https://doi.org/10.1016/j.aap.2021.106281.
[6] V.V. Dixit, S. Chand, D.J. Nair, Autonomous vehicles: disengagements, accidents and reaction times, PLoS One 11 (12) (2016) e0168054, https://doi.org/10.1371/journal.pone.0168054.
[7] F.M. Favarò, N. Nader, S.O. Eurich, M. Tripp, N. Varadaraju, Examining accident reports involving AVs in California, PLoS One 12 (9) (2017) e0184952, https://doi.org/10.1371/journal.pone.0184952.
[8] S.S. Banerjee, S. Jha, J. Cyriac, Z.T. Kalbarczyk, R.K. Iyer, Hands off the wheel in autonomous vehicles?: a systems perspective on over a million miles of field data, June, in: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), IEEE, 2018, pp. 586–597, https://doi.org/10.1109/DSN.2018.00066.
[9] S.H. Leilabadi, S. Schmidt, In-depth analysis of autonomous vehicle collisions in California, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE, 2019, October, pp. 889–893, https://doi.org/10.1109/ITSC.2019.8916775.
[10] S. Wang, Z. Li, Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data, Accid. Anal. Prev. 129 (2019) 44–54, https://doi.org/10.1016/j.aap.2019.04.015.
[11] H. Alambeigi, A.D. McDonald, S.R. Tankasala, Crash Themes in Automated Vehicles: A Topic Modeling Analysis of the California Department of Motor Vehicles Automated Vehicle Crash Database, 2020, https://doi.org/10.48550/arXiv.2001.11087 arXiv preprint arXiv:2001.11087.
[12] ISO, 26262 – Road Vehicles – Functional Safety, 2018.
[13] PEGASUS Project Consortium, PEGASUS method: an overview, Available: https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUSGesamtmethode.pdf, 2019.
[14] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, F. Diermeyer, Survey on scenario-based safety assessment of automated vehicles, IEEE Access 8 (2020) 87456–87477, https://doi.org/10.1109/ACCESS.2020.2993730.
[15] M. Steimle, T. Menzel, M. Maurer, Toward a consistent taxonomy for scenario-based development and test approaches for automated vehicles: a proposal for a structuring framework, a basic vocabulary, and its application, IEEE Access 9 (2021) 147828–147854, https://doi.org/10.1109/ACCESS.2021.3123504.
[16] M. Scholtes, L. Westhofen, L.R. Turner, K. Lotto, M. Schuldes, H. Weber, L. Eckstein, 6-layer model for a structured description and categorization of urban traffic and environment, IEEE Access 9 (2021) 59131–59147, https://doi.org/10.1109/ACCESS.2021.3072739.
[17] J.E. Park, W. Byun, Y. Kim, H. Ahn, D.K. Shin, The impact of automated vehicles on traffic flow and road capacity on urban road networks, J. Adv. Transport. (2021), https://doi.org/10.1155/2021/8404951, 2021.
[18] Y. Song, M.V. Chitturi, D.A. Noyce, Automated vehicle crash sequences: patterns and potential uses in safety testing, Accid. Anal. Prev. 153 (2021) 106017, https://doi.org/10.1016/j.aap.2021.106017.
[19] S. Shanthi, R.G. Ramani, Feature relevance analysis and classification of road traffic accident data through data mining techniques, Proceedings of the World Congress on Engineering and Computer Science 1 (2012, October) 24–26, sn.
[20] M. Taamneh, S. Alkheder, S. Taamneh, Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates, J. Transport. Saf. Secur. 9 (2) (2017) 146–166, https://doi.org/10.1080/19439962.2016.1152338.
[21] L.J. Muhammad, S. Sani, A. Yakubu, M.M. Yusuf, T.A. Elrufai, I.A. Mohammed, A.M. Nuhu, Using decision tree data mining algorithm to predict causes of road traffic accidents, its prone locations and time along Kano–Wudil highway, International Journal of Database Theory and Application 10 (1) (2017) 197–206.
[22] T.K. Bahiru, D.K. Singh, E.A. Tessfaw, Comparative study on data mining classification algorithms for predicting road traffic accident severity, April, in: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, 2018, pp. 1655–1660, https://doi.org/10.1109/ICICCT.2018.8473265.
[23] G. Janani, N.R. Devi, Road traffic accidents analysis using data mining techniques, JITA-JOURNAL OF INFORMATION TECHNOLOGY AND APLICATIONS 14 (2) (2016), https://doi.org/10.7251/JIT1702084J.
[24] L. Li, S. Shrestha, G. Hu, Analysis of road traffic fatal accidents using data mining techniques, June), in: 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE, 2017, pp. 363–370, https://doi.org/10.1109/SERA.2017.7965753.

[25] A.M. Boggs, B. Wali, A.J. Khattak, Exploratory analysis of automated vehicle crashes in California: a text analytics & hierarchical Bayesian heterogeneity-based approach, Accid. Anal. Prev. 135 (2020) 105354, https://doi.org/10.1016/j.aap.2019.105354.
[26] S. Lee, R. Arvin, A.J. Khattak, Advancing investigation of automated vehicle crashes using text analytics of crash narratives and Bayesian analysis, Accid. Anal. Prev. 181 (2023) 106932, https://doi.org/10.1016/j.aap.2022.106932.
[27] H. DrissiTouzani, S. Faquir, A. Yahyaouy, Data mining techniques to analyze traffic accidents data: case application in Morocco, October), in: 2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS), IEEE, 2020, pp. 1–4, https://doi.org/10.1109/ICDS50568.2020.9268729.
[28] L. Lin, Q. Wang, A.W. Sadek, Data mining and complex network algorithms for traffic accident analysis, Transport. Res. Rec. 2460 (1) (2014) 128–136, https://doi.org/10.3141/2460-14.
[29] S. Kumar, D. Toshniwal, A data mining approach to characterize road accident locations, Journal of Modern Transportation 24 (1) (2016) 62–72, https://doi.org/10.1007/s40534-016-0095-5.
[30] X. Kong, S. Das, Y. Zhang, L. Wu, J. Wallis, In-depth understanding of near-crash events through pattern recognition, Transport. Res. Rec. 2676 (12) (2022) 775–785, https://doi.org/10.1177/03611981221097395.
[31] X. Wang, Y. Peng, T. Xu, Q. Xu, X. Wu, G. Xiang, H. Wang, Autonomous driving testing scenario generation based on in-depth vehicle-to-powered two-wheeler crash data in China, Accid. Anal. Prev. 176 (2022) 106812, https://doi.org/10.1016/j.aap.2022.106812.
[32] P. Nitsche, P. Thomas, R. Stuetz, R. Welsh, Pre-crash scenarios at road junctions: a clustering method for car crash data, Accid. Anal. Prev. 107 (2017) 137–151, https://doi.org/10.1016/j.aap.2017.07.011.
[33] B. Sui, N. Lubbe, J. Bärgman, A clustering approach to developing car-to-two-wheeler test scenarios for the assessment of Automated Emergency Braking in China using in-depth Chinese crash data, Accid. Anal. Prev. 132 (2019) 105242, https://doi.org/10.1016/j.aap.2019.07.018.
[34] Q. Yuan, X. Xu, J. Zhau, Paving the way for autonomous vehicle testing in accident scenario analysis of yizhuang development Zone in Beijing, in: CICTP 2020, 2020, pp. 62–72.
[35] X. Kong, S. Das, Y. Zhang, Mining patterns of near-crash events with and without secondary tasks, Accid. Anal. Prev. 157 (2021) 106162, https://doi.org/10.1016/j.aap.2021.106162.
[36] D. Pan, Y. Han, Q. Jin, H. Wu, H. Huang, Study of typical electric two-wheelers pre-crash scenarios using K-medoids clustering methodology based on video recordings in China, Accid. Anal. Prev. 160 (2021) 106320, https://doi.org/10.1016/j.aap.2021.106320.
[37] Z. Tan, Y. Che, L. Xiao, W. Hu, P. Li, J. Xu, Research of fatal car-to-pedestrian precrash scenarios for the testing of the active safety system in China, Accid. Anal. Prev. 150 (2021) 105857, https://doi.org/10.1016/j.aap.2020.105857.
[38] E. Esenturk, D. Turley, A. Wallace, S. Khastgir, P. Jennings, A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles, International Journal of Transportation Science and Technology (2022), https://doi.org/10.1016/j.ijtst.2022.10.002.
[39] M. Kang, W. Lee, K. Hwang, Y. Yoon, Vision transformer for detecting critical situations and extracting functional scenario for automated vehicle safety assessment, Sustainability 14 (15) (2022) 9680, https://doi.org/10.3390/su14159680.
[40] N. Novat, E. Kidando, B. Kutela, A.E. Kitali, A comparative study of collision types between automated and conventional vehicles using Bayesian probabilistic inferences, J. Saf. Res. 84 (2023) 251–260, https://doi.org/10.1016/j.jsr.2022.11.001.
[41] S. Das, A. Dutta, I. Tsapakis, Automated vehicle collisions in California: applying Bayesian latent class model, IATSS Res. 44 (4) (2020) 300–308, https://doi.org/10.1016/j.iatssr.2020.03.001.
[42] J. Torres, Y. Li, J. Zhang, Investigating traffic crashes involving autonomous vehicles, in: IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE), 2021, pp. 1046–1051.
[43] M.T. Ashraf, K. Dey, S. Mishra, M.T. Rahman, Extracting rules from autonomous-vehicle-involved crashes by applying decision tree and association rule methods, Transport. Res. Rec. 2675 (11) (2021) 522–533, https://doi.org/10.1177/03611981211018461.
[44] M. Kang, J. Song, K. Hwang, The extraction of automated vehicles traffic accident factors and scenarios using real-world data, July), in: Congress on Intelligent Systems: Proceedings of CIS 2021, ume 1, Springer Nature Singapore, Singapore, 2022, pp. 1–15, https://doi.org/10.1007/978-981-16-9416-5_1.
[45] C. Stark, C. Medrano-Berumen, M.İ. Akbaş, Generation of autonomous vehicle validation scenarios using crash data, March, in: 2020 SoutheastCon, IEEE, 2020, pp. 1–6, https://doi.org/10.1109/SoutheastCon44009.2020.9249662.
[46] https://www.dmv.ca.gov.
[47] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (1967) 241–254, https://doi.org/10.1007/BF02289588.
[48] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 2 (1) (2012) 86–97, https://doi.org/10.1002/widm.53.
[49] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, September, in: Proc. 20th Int. Conf. Very Large Data Bases, vol. 1215, VLDB, 1994, pp. 487–499.
[50] ENABLE-S3 Project Consortium, Testing and Validation of Highly Automated Systems: Summary of Results, 2019 [Online]. Available: https://drive.google.com/?le/d/15c1Oe69dpvW5dma8-uS8hev17x-6V3zU/view.
[51] E. de Gelder, O.O. den Camp, N. de Boer, Scenario Categories for the Assessment of Automated Vehicles, Version, CETRAN, Singapore, 2020, p. 1.
[52] W.G. Najm, J.D. Smith, M. Yanagisawa, Pre-crash Scenario Typology for Crash Avoidance Research (No. DOT-VNTSC-NHTSA-06-02), National Highway Traffic Safety Administration, United States, 2007.
[53] W. Ko, S. Park, J. Yun, S. Park, I. Yun, Development of a framework for generating driving safety assessment scenarios for automated vehicles, Sensors 22 (2022) 6031, https://doi.org/10.3390/s22166031.