A Comparison of Patterns and Contributing Factors of ADAS and ADS involved Crashes

Song Yan[a], Chunxi Huang[b], Dengbo He[a,c,d*]

[a] *Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangdong, China;* [b] *Interdisciplinary Programs Office (IPO), The Hong Kong University of Science and Technology, Hong Kong SAR, China;* [c] *HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China;* [d] *Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China*

*Corresponding author. Email: dengbohe@ust.hk

Song YAN received his bachelor's degree in automotive engineering from Zhejiang University, China, and master's degree in mechanical engineering from the University of Tokyo, Japan. He is currently a PhD student in Intelligent Transportation Thrust at the Hongkong University of Science and Technology (Guangzhou). His research interests include automated driving systems, human factors, and driver behaviors.

Chunxi HUANG is currently a Ph.D. student in Robotics and Autonomous Systems program at The Hong Kong University of Science and Technology, Hong Kong. Before that, he obtained his M.S. degree in Industrial & Systems Engineering (2020) and Bachelor's degree in Industrial Engineering (2018) from Korea Advanced Institute of Science and Technology, Daejeon, South Korea and Zhejiang University, Hangzhou, China, respectively.

Dengbo HE is an Assistant Professor from the Systems Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He is also with the Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong SAR, China. He received his Ph.D. degree from the University of Toronto, Canada in 2020, his M.S. Degree from the Shanghai Jiao Tong University, China, in 2016, and his Bachelor's degree from the Hunan University, China in 2012.

A Comparison of Patterns and Contributing Factors of ADAS and ADS involved Crashes

Crashes involving Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS) have been increasing in recent years. Understanding the characteristics of these crashes can guide the optimization of driving automation systems and the policies improving the safety of mixed traffic. However, due to the limited available data, the crashes of ADS- and ADAS-controlled vehicles are still under-investigated. Thus, utilizing the latest National Highway Traffic Safety Administration crash reports, our study explores the patterns and contributing factors of ADAS- and ADS-involved crashes. The sequences of events leading to crashes were extracted from the reports and then categorized into five clusters. Next, for incomplete records, a non-parametric imputation method was applied based on Random Forest. Finally, logistic regression models were built to explore the factors associated with the crashes. The results show that the automation level, speed limit, and vehicle speed are predictors of crash patterns. At the same time, the crash pattern, combined with incident time, roadway type, roadway surface, and vehicle model year are associated with crash outcomes (i.e., contact area and injury severity). The results indicate that further improvement of the ADS/ADAS control algorithms and driver education, may be needed to improve the safety of mixed traffic.

Keywords: ADAS; ADS; sequence analysis; clustering analysis.

1. Introduction

With the advancement of automated driving technologies, the past few years have witnessed a rapid implementation of Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS) on the road. In 2020, 50% of new car models in the market are provided with Society of Automotive Engineers (SAE) Level 2 systems (e.g., adaptive control, lane keeping systems, etc.) (SAE International, 2021) as standard or optional features (Consumer Reports, 2021), and this ratio is expected to surpass 90% in the 2040s (Highway Loss Data Institute, 2022). It is believed that ADAS and ADS can improve traffic safety by reducing the number of accidents caused by human errors

(Litman, 2022). However, crashes involving these systems still happen. Thus, it is essential to understand the patterns of the ADS- and ADAS-involved crashes and the factors leading to the crashes so that further countermeasures can be adopted to reduce the crash rates of ADS- and ADAS-controlled vehicles (e.g., designing testing scenarios for ADS- and ADAS-controlled vehicles and training human drivers to better share the road with driving automation).

Extensive studies have been conducted on crash characterization, aiming at identifying contributing factors of ADS- and ADAS-involved crashes and proposing and validating crash mitigation countermeasures. The majority of these studies characterized crashes in the conventional approach which only considered the static pre-crash information that can be directly extracted from the crash reports (e.g., weather condition, lighting condition, roadway type, and vehicle motion) and classified the crashes into general post-crash types, such as head-on, sideswipe, rear-end, etc (Chen et al., 2020; Dadvar & Ahmed, 2021; Esenturk et al., 2021; Kutela et al., 2022). A few other studies adopted a more nuanced approach which delved into the chronological progression of the crashes. They explored the patterns of crashes by extracting and analyzing the sequence of events derived from text narratives (Song et al., 2021, 2022; Wu et al., 2016). However, few studies have taken both the static pre-crash variables and the crash progression information into consideration in crash characterization and thus we have little knowledge of how static factors (e.g., environmental condition) and traffic dynamics (i.e., progression information) jointly influence the ADS- and ADAS-involved crashes.

Additionally, crash characterization heavily relies on the data sources. The CA DMV crash dataset (California DMV, 2022) has been mostly adopted in previous research, but it only covers ADS-involved crashes, while the crashes involving ADAS remain conspicuously underexplored due to the absence of readily accessible ADAS-

specific crash datasets. Consequently, our understanding of the distinctive crash patterns and contributing factors associated with ADAS-involved incidents remains limited. Previous research has pointed out that, on public roads, drivers may adopt different strategies when interacting with vehicles with different levels of driving automation (Huang, Wen, He, et al., 2022; Wen et al., 2023). Considering the inherent difference between ADAS and ADS (i.e., whether human intervention is required), it is necessary to distinguish between ADAS and ADS to better understand the safety implications of driving automation technologies and develop appropriate crash mitigation measures.

Thus, through a comprehensive analysis of the latest dataset that involves both ADAS and ADS crashes, our study seeks to identify and compare the patterns and contributing factors of ADAS and ADS-involved crashes in order to improve ADAS and ADS safety. The contributions of this study are as follows: First, the study identifies the crash patterns based on crash sequences extracted from narratives and proposes a framework for modeling crash outcomes, in which the crash pattern is considered an intermediate factor. Second, our study shows that in combination with other contributing factors, the crash pattern can influence crash outcomes (i.e., contact area and injury severity). Finally, the results of our study highlight the importance of distinguishing ADAS and ADS in crashes when analyzing the factors contributing to crashes.

2. Literature Review

2.1 Crash datasets

To facilitate traffic safety research and reduce traffic incidents, authoritative organizations around the world have been working on collecting and publishing vehicle crash datasets in the past few decades. The National Automotive Sampling System General Estimates System (NASS-GES) database and the Fatality Analysis Reporting System (FARS) published by the National Highway Traffic Safety Administration

(NHTSA) are the most commonly adopted datasets in traffic safety research (NHTSA, 2023). The NASS-GES database, starting in 1988, is a comprehensive database comprising police-reported incidents encompassing various road users, serving various purposes such as assessing overall crash trends, pinpointing safety issues on roads, and evaluating crash mitigation strategies. The FARS system, on the other hand, primarily focuses on fatal traffic crashes that result in the death of the traffic participants. The major limitation of these datasets is that they only included crashes involving human-driven vehicles, which cannot fully reveal the characteristics of ADAS and ADS crashes, and thus are not suitable for ADAS or ADS-related research.

With the SAE Level-2 vehicles gradually becoming available in the market and SAE Level-3 or higher-level vehicles being tested on public roads, the number of ADS-involved and ADAS-involved crashes has been increasing in recent years. Since 2014, the CA DMV has required automotive companies to report the crash and disengagement cases of ADS tested on California roads (California DMV, 2022). This dataset contains information regarding ADS-involved crashes, including the environmental conditions, vehicle information, and crash outcomes. The CA DMV dataset was the only publicly available driving-automation-related crash dataset since it was published. However, the CA DMV does not contain ADAS-involved crashes. In 2022, NHTAS released a new dataset that contains both ADAS and ADS crash reports (NHTSA, 2022), which enables in-depth analysis of the difference between ADS crash patterns and ADAS patterns.

2.2 Research on ADS-involved crashes

Using the CA DMV dataset, a variety of methods have been applied to explore the contributing factors of crashes, such as logistic regression (Esenturk et al., 2021), classification tree (Dadvar & Ahmed, 2021), XGBoost (Chen et al., 2020), and Bayesian

networks (Kutela et al., 2022). They found that the vehicle movement (e.g., proceeding straight or making a turn), lighting condition, road surface condition, road type, incident time, crash location, and the speed limit at the site were correlated with the rate and outcomes of ADS-involved crashes (Das et al., 2020; Kutela et al., 2022; Ren et al., 2022; Torres et al., 2021; Xu et al., 2019). Further, through in-depth comparative analysis, Liu et al. (2021) explored the difference between ADS-involved crashes and conventional vehicle crashes in terms of leading factors and crash characteristics.

However, Wu et al. (2016) argued that the conventional practices, which modeled the crashes based on static pre-crash and post-crash factors, have ignored the chronological progression of crashes. In other words, although various methods have been proposed and proven effective in identifying crash patterns and contributing factors of ADS-involved crashes in existing studies, the information contained in the crash reports was not fully exploited to characterize the crashes. Thus, the authors suggested that the sequence of events extracted from report narratives should be used for crash characterization, based on which several types of crash sequences associated with severe crashes can be identified through clustering analysis. This method has been adopted by Song et al. (2021) for the analysis of crashes in the CA DMV dataset, and they identified seven crash patterns of ADS-controlled vehicles. Later on, Song et al. (2022) further investigated the correlations among crash sequence, crash outcomes, environmental conditions, and human-related factors of crash partners (i.e., the characteristics of the human-driven vehicles involved in crashes, for example, speeding, careless driving, and improper operations) using Bayesian network modeling. However, it remains to be explored how the pre-crash factors affect the progression of crashes and finally lead to different crash outcomes of ADS-controlled vehicles.

2.3 Research on ADAS-involved crashes

Though some studies have investigated the causes and characteristics of HDV crashes and discussed the potential benefits and challenges of ADAS in preventing these crashes (Galloway et al., 2023; Scanlon et al., 2015, 2016; Seacrist et al., 2021), the amount of research on ADAS-involved crashes is limited due to the lack of publicly available dataset. For example, based on the SHRP 2 naturalistic driving dataset, Seacrist et al. (2021) analyzed critical driver errors (e.g., distractions, decision errors, performance errors, etc.) that contribute to crashes among different age groups, and they pointed out that the ADAS functions (e.g., forward collision warning, high-speed warning, automatic emergency braking, etc.) has the potential to reduce driver-related errors. Similarly, based on the National Motor Vehicle Crash Causation Survey (NMVCCS) database, Scanlon et al. (2016) investigated the pre-crash kinematics of crashes happening at intersections and discussed the possibility of using ADAS to prevent crashes.

However, it is important to emphasize that the current understanding of ADAS-involved crashes is insufficient, as mentioned, mainly due to the lack of available datasets. The release of the NHTSA crash dataset in 2022 has made analysis of ADAS-involved crashes possible (Ding et al., 2023). Some studies have compared the NHTSA dataset to existing traffic incident datasets (Goodall, 2023) and even strived to build a unified dataset encompassing all types of vehicle crashes (Zheng et al., 2023). A few studies have already analyzed the characteristics of specific types of ADAS-involved crashes using the NHTSA dataset (e.g., Huang, Wen, & He, 2022). More efforts, however, are still needed to understand the factors leading to ADAS-involved crashes in order to design countermeasures to improve the safety of ADAS-equipped vehicles.

3 Data Preparation

NHTSA has required automobile manufacturers to report crash cases involving ADAS and ADS systems since Jul. 2021. Till Nov. 2022, a total of 1374 ADAS crash reports, and 475 ADS crash reports were recorded (NHTSA, 2022). In the NHTSA crash reports, ADS refers to the automated driving systems with SAE Level 3 or higher-level automation, which can "*perform the entire dynamic driving task on a sustained basis within a defined operational design domain without driver involvement*"; ADAS refers to the SAE Level 2 systems which "*provide both speed and steering input when the driver assistance system is engaged but require the human driver to remain fully engaged in the driving task at all times*". The data screening process is shown in **Error! Reference source not found.**

In the original dataset, there might be multiple versions of reports for one crash, which can be identified with 'Report ID' and 'Report Version' in the reports. Therefore, in the first step, the duplicates were removed using a Python script (based on Python 3.7 and Pandas 1.3.4 library), and the latest report of a crash was kept. In Step 2, for 555 out of 697 ADS-involved crash reports and 80 out of 357 ADAS-involved crash reports, the narrative is protected from disclosure due to business concerns, which can be recognized as 'Narrative_CBI'. These reports were also removed using the Python script. Then, in Step 3, the dataset went through a manual screening to eliminate duplicates that cannot be identified with 'Report ID' and 'Report Version', and the reports that do not contain a sufficient description of the accident progression in their narratives. Finally, we obtained 92 valid ADAS crash reports and 100 valid ADS crash reports, and the subsequent analysis was conducted based on these reports. It is worth noting that, as the reports were recorded manually and relied on self-reporting from vehicle owners or operating entities, a large number of reports were abandoned due to duplicates and incomplete or

confidential information in the reports. Specifically, a larger portion of ADAS crash data (93.3%) was filtered out in the screening process compared to that of the ADS crash data (75.8%). This is because ADAS-equipped vehicles are primarily owned by consumers, making the complete collection of the data more difficult. In contrast, ADS-equipped vehicles are owned by manufacturers, allowing for more complete, timely, and standard crash reports.

Table 1. Data screening process.

Screening Process	**Num. of valid reports**	
	ADAS	**ADS**
Raw data (Latest update: Nov. 15, 2022)	1374	475
Step 1: Remove duplicates (Auto filtering based on 'Report ID' and 'Report Version')	697	357
Step 2: Remove Confidential Business Information (CBI) (Auto filtering based on 'Narrative_CBI')	142	277
Step 3: Manual selection (Manual selection based on 'Narrative')	92	110

4 Methodology

Error! Reference source not found. provides an overview of the methodology adopted in this study. The NHTSA crash reports consist of two types of data, i.e., the narratives and the structured crash data (i.e., static information regarding the scenario where the crash happened, such as incident time, weather, and speed of the involved road agents). The narratives in the crash reports contain information regarding the chronological development of crashes that cannot be revealed in the structural data of the reports. To extract useful information from the narratives, we followed the sequence analysis method proposed by Wu et al. (2016), which has been commonly used to extract crash sequences and identify crash patterns from narratives. Specifically, we encoded the narratives into crash sequences (step 1) and then calculated the distance between each pair of sequences

(step 2), based on which we performed clustering analysis and found several crash patterns (step 3). To eliminate the disturbance of unbalanced data and missing values, we performed data aggregation and imputation (step 4). Finally, using the crash patterns and structured crash data, we built logistic regression models and identified factors that influence crash patterns and crash outcomes (i.e., contact area and injury severity). In the following sections, we will explain the above steps in more detail.

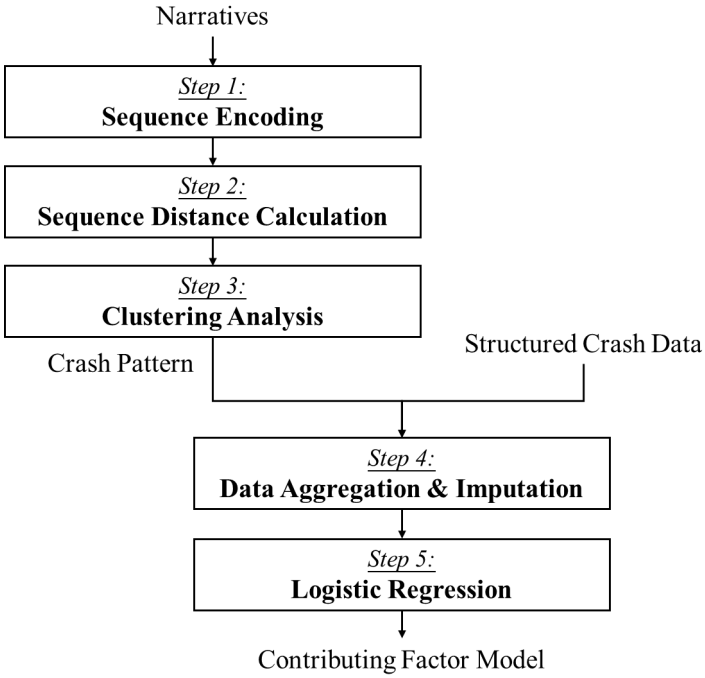

Figure 1. Methodology flow chart

4.1 Sequence encoding

To make full use of the crash data, the crash details in the narratives were manually extracted and encoded into event sequences according to the encoding scheme shown in Table 2. The encoding scheme was mainly adopted from the method proposed by Song et al. (Song et al., 2021), and some modifications were made to adapt the narratives of the NHTSA dataset, i.e., the codes corresponding to nonexisting events in the narratives were removed, and new events (including E1, E2, DT, and F) were newly added. In the study of Song et al. (2021), “merge left” and “merge right” were assigned with different

codes. However, in our study, the two maneuvers were encoded into a single code (M1 or M2) since the direction of merging maneuvers is not mentioned in most narratives. To facilitate subsequent analysis, we categorized the events into pre-crash and crash (Table 2) so that we can distinguish between inter-group mismatch and intra-group mismatch when comparing two events, which allows a more refined way of measuring the dissimilarity between sequences.

Table 2. Sequence encoding scheme.

Code	Description	Code	Description
	Pre-crash events		
A1	v1 accelerate/proceed	M2	v2 merge left/right
A2	v2 accelerate/proceed	R1	v1 turn right
B2	v2 back up	S1	v1 stop
D1	v1 decelerate	S2	v2 stop
E1	v1 entering traffic	TO	v1 driver takes over the driving task
E2	v2 entering traffic	DT	v1 driver distraction / wrong operation
L1	v1 turn left	F	v1 disengage / misfunction
L2	v2 turn right	V1	v1 violate the rule
M1	v1 merge left/right	V2	v2 violate the rule
	Crash events		
X10	v1 hit object	X13	v1 contact v3
X12	v1 contact v2	X23	v2 contact v3

Note: in the table, v1 is the subject vehicle (SV) equipped with ADAS or ADS; v2 refers to a second-party crash partner (CP) involved in the crash; and v3 indicates a third-party CP involved in the crash. For readability, we used the codes in the left columns to represent the events described in the right columns. The codes of events are later linked to form the crash sequence.

The following is an example of encoding a narrative into a crash sequence. The original narrative in NHTSA crash reports is "*The customer was driving straight on a highway at approximately 66 miles per hour when the wheels allegedly locked up, causing the customer to lose control and hit the concrete barrier.*" (NHTSA, 2022). Based on the method mentioned above, we encoded the narrative into a crash sequence as "A1-F-X10".

4.2 Sequence distance calculation

To quantitatively describe the difference between crashes, we calculated the distance between each pair of sequences. The Needleman-Wunsch (NW) algorithm (Needleman & Wunsch, 1970) and the Dynamic Time Warping (DTW) algorithm (Arribas-Gil & Müller, 2014) are the most commonly used methods for global optimal alignment. When aligning two sequences, the DTW algorithm stretches the sequences and fills in the missing part with an adjacent event; in contrast, the NW algorithm fills the missing part with gaps (Huang et al., 2019). The pairing results obtained through the DTW algorithm may either match or mismatch, while the NW algorithm can produce matches, mismatches, or gaps. Each type of pairing result can be assigned a distinct score. Thus, the NW allows more complicated algorithm design. Given that the NW algorithm has also been proven to be effective in crash sequence analysis in previous studies (Song et al., 2021, 2022; Wu et al., 2016) and compared to DTW, the proximity matrix from the NW led to categories with more distinct characteristics in the following cluster analysis in our study, the NW algorithm was adopted.

Figure 2 is an example of aligning two crash sequences using the NW algorithm, where the elements from two sequences are compared in pairs. As mentioned, the pairing result between two elements can be a match, mismatch, or gap.

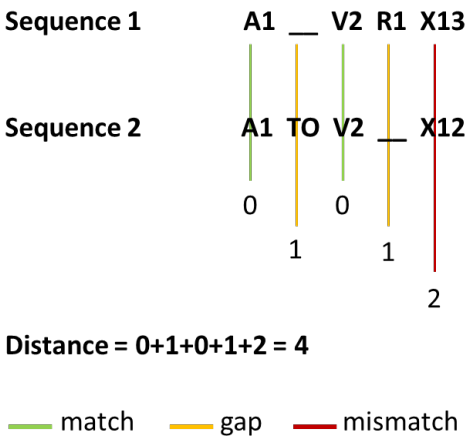

Figure 2. An example of crash sequence alignment

Different scores are assigned to different pairing results based on the scoring system given in Table 3. In the equations, e1 indicates an element of sequence 1, e2 refers to an element of sequence 2, and the underscore symbol represents inserting a gap in the sequence. It is worth noting that we have divided the elements into the pre-crash events and the crash events in the encoding scheme. Therefore, in the scorning system, we need to consider two types of mismatches: if two different elements are the same type of event, the pairing result is an intragroup mismatch, while if the two elements are different types of events, the pairing result is an intergroup mismatch. The pairing score measures the distance between two elements; the higher the score, the larger the distance. Finally, the overall distance of two aligned sequences is given by the sum of the pairing scores.

Table 3. Scoring system.

Pairing result	**Score**
Gap	$s(\_, e2) = s(e1, \_) = 1$
Match	$s(e1, e2) = 0$
Mismatch	$s(e1, e2) = \begin{cases} 2, \ intragroup \\ 3, \ intergroup \end{cases}$

In order to find the optimal alignment that yields the smallest distance, the NW algorithm constructs an n*m matrix ($n = 1 + size\ of\ sequence1$, $m = 1 + size\ of\ sequence2$) as shown in Figure 3.

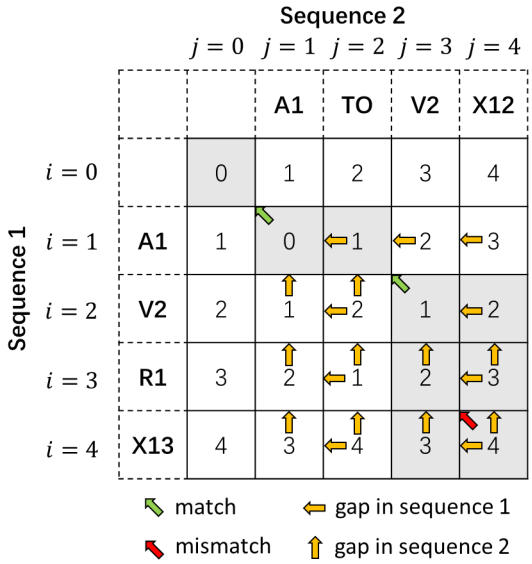

Figure 3. An example of the alignment matrix

The alignment starts at the top left corner and ends at the bottom right corner. The value of each cell is given by equation (1):

$$f(i,j) = \begin{cases} i, & j = 0 \\ j, & i = 0 \\ min \begin{cases} f(i-1,j-1) + s(seq1[i], seq2[j]) \\ f(i-1,j) + s(seq1[i], \_) \\ f(i,j-1) + s(\_, seq2[j]) \end{cases} & , \ i,j \neq 0 \end{cases} \quad (1)$$

where $seq1[i]$ indicates the ith element in sequence 1, and $seq2[j]$ indicates the jth element of sequence 2. The function $s()$ is used to calculate the pairing scores according to the scoring system in Table 3. The value of the bottom right cell is the smallest overall distance between the two sequences. The optimal alignments can be found by tracing back from the bottom right corner to the top left corner. As shown in Figure 3, the smallest sequence distance of this example is 4, and there are four optimal alignments (see Table 4).

Table 4. Optimal alignments of the example.

	Optimal alignment 1	**Optimal alignment 2**	**Optimal alignment 3**	**Optimal alignment 4**
Sequence 1	A1-__-V2-__-R1-X13	A1-__-V2-R1-__-X13	A1-__-V2-R1-X13	A1-__-V2-R1-X13-__

Sequence 2 A1-TO-V2-X12-__-__A1-TO-V2-__-X12-__ A1-TO-V2-__-X12 A1-TO-V2-__-__-X12

4.3 Clustering analysis

We calculated the distance between each pair of crash sequences to generate the distance matrix, which is used as the input for clustering analysis. The total number of crash sequences is 202 (including 92 ADAS crashes and 110 ADS crashes), and the size of the distance matrix is 202 by 202. We compared the performance of several unsupervised machine learning methods, including k-means (MacQueen, 1967), agglomerative hierarchical clustering (Müllner, 2011), and density-based clustering algorithm (Ester et al., 1996). It turns out that the agglomerative hierarchical clustering algorithm with a bottom-up manner and a complete linkage yielded the best clustering performance and thus it was adopted to group the crash sequences into clusters.

The silhouette width was used to evaluate the clustering performance and determine the most appropriate number of clusters, which measures how well a sample (i.e., crash sequence) matches its own cluster in comparison with other clusters (Rousseeuw, 1987). The silhouette width of a sample is given by:

$$s[i] = \frac{b[i]-a[i]}{\max(a[i],b[i])} \quad (2)$$

where $s[i]$ is the silhouette width of the sample i, $b[i]$ is the average distance of sample i to the samples in its own cluster, and $a[i]$ is the average distance of sample i to the samples in its nearest cluster. The average silhouette width (ASW) of all the samples is given by:

$$ASW = \frac{1}{N}\sum_{i=1}^{N} s[i] \quad (3)$$

where N is the total number of samples.

4.4 Data aggregation and imputation

The 202 valid crash reports used in this study consisted of 92 ADAS cases and 110 ADS cases. However, of these reports, only 4 ADAS cases and 44 ADS cases were completely recorded (i.e., without missing values). Considering the limited number of usable crash reports, and even smaller number of completely recorded cases, it is necessary to impute the missing values to overcome the issue of small sample size and enable follow-up regression analysis. Further, given that many of the categorical variables are over-divided and the dataset is relatively imbalanced with too few samples in some minor classes (e.g., only one ADS-involved crash resulted in severe injury), we aggregated some categorical variables to avoid quasi-complete or complete separation of data points in the maximum likelihood estimates process of logistic regression analysis. Table 5 shows how the data was aggregated. Specifically, we have merged the model year into 'Prio 2020' and '2020 or later' and the incident time into 'Day' and 'Night' to ensure a balanced distribution of samples. Following Kaber et al. (2012), the roadway type has been categorized into 'Simple' and 'Complex' based on the complexity of the roadway layout. Compared to simple roadways, complex roadways involve more conflict points (location at which traffic movements intersect such as crossing, merging, and diverging) that are associated with high crash risks (Lu et al., 2013). The SV contact area has been categorized as 'longitudinal' and 'non-longitudinal' (i.e., lateral or vertical) as this is often how vehicle motion is decoupled in the design of motion control and collision avoidance functions (Cheng et al., 2020; Zhang et al., 2022).

Table 5. Data aggregation.

Variable	Original data	Aggregated data
Incident time	Numerical (HH:mm)	• Day (6:00~16:59) • Night (17:00~5:59)

Model year	Numerical (2015~2023)	• Prior 2020 • 2020 or later
Weather	• Clear • Snow, Cloudy, Fog/Smoke, Rain, Severe Wind	• Clear • Bad
Roadway surface	• Dry • Wet, Snow/Slush/Ice	• Dry • Wet
Roadway type	• Street, Highway/Freeway, Rural Road • Intersection/ Parking Lot/ Traffic Circle	• Simple • Complex
SV contact area	• Involves front, rear, or both • Does not involve front or rear	• Longitudinal • Non-Longitudinal
Highest injury severity	• No injuries reported • Minor, Moderate, and Severe	• No injury • With injury

Note: In the dataset, following the definition of NHTSA, 10 contact areas are defined, including front, front left, front right, rear, rear left, rear right, left, right, top, and bottom. The subject vehicle can be contacted at a single area or multiple areas during the crashes.

Figure 4 provides a visualization of the data completeness of each variable. The height of each bar indicates the number of observed samples of the corresponding variable. Among the 12 variables used in this study, three variables are completely observed, namely automation level (ADS vs. ADAS), model year, and crash pattern (from cluster analysis). The remaining variables with missing values were imputed using the MissForest algorithm, a non-parametric imputation method based on Random Forest. MissForest has been widely used in biology, medicine, and machine learning due to its low imputation error and ability to handle mixed-type data (i.e., data with both categorical and continuous variables) (Stekhoven & Buhlmann, 2012).

To validate the imputation accuracy, we calculated the Jensen-Shannon (JS) Divergence (Lin, 1991) between the originally observed data and the imputed data. The JS Divergence is a measure of the distributional dissimilarity between two samples, which is given by equation (4) (for discrete variables) and equation (5) (for continuous variables):

$$JS_{discrete}(p||q) = \frac{1}{2}\sum_{i=1}^{n} p(x_i)\log\left(\frac{2p(x_i)}{p(x_i)+q(x_i)}\right) + \frac{1}{2}\sum_{i=1}^{n} q(x_i)\log\left(\frac{2q(x_i)}{p(x_i)+q(x_i)}\right) \quad (4)$$

$$JS_{continuous}(p||q) = \frac{1}{2}\int p(x)\log\left(\frac{2p(x)}{p(x)+q(x)}\right)dx + \frac{1}{2}\int p(x)\log\left(\frac{2q(x)}{p(x)+q(x)}\right)dx \quad (5)$$

where p is the distribution of the observed samples, and q is the distribution of all samples, including the imputed ones.

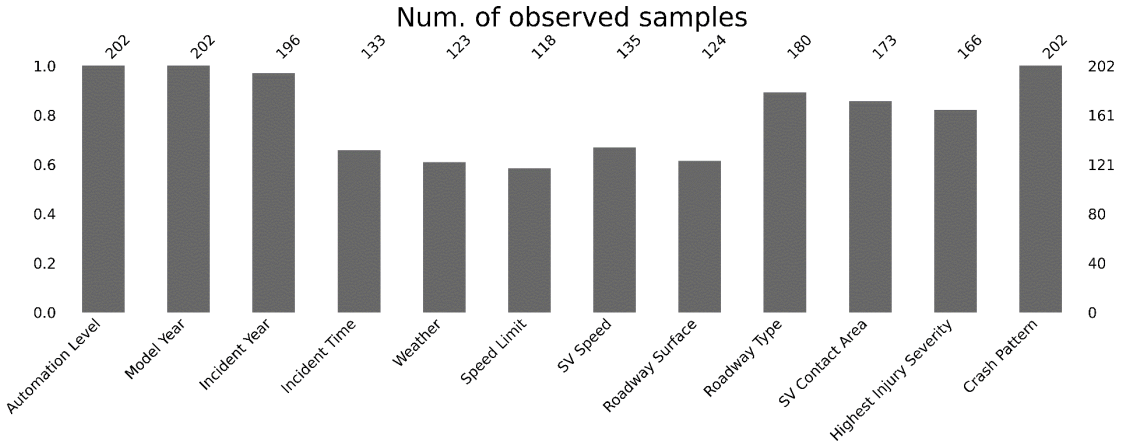

Figure 4. Data completeness of each variable (total sample: 202).

4.5 Logistic Regression

Finally, a logistic regression analysis was performed to determine how the crash pattern works in conjunction with other factors to influence crash outcomes. Compared to the contributing factor models commonly used in previous studies, we introduced several new influential factors (i.e., automation level and crash pattern) into the model. Here we have classified the variable into static factors, crash progression, and crash outcomes (see Figure 5). We hypothesize that some factors have a direct effect on the crash outcome, while others influence the outcome by acting on the crash progression. Therefore, three logistic regression models were fitted using the SAS LOGISTIC process based on the hypothesized contributing factor model shown in Figure 5.

In model 1, the crash pattern was used as the dependent variable, and the static factors were used as the independent variables in order to investigate how the static factors contribute to the crash progression. In model 2 and model 3, SV contact area and highest injury severity were considered as dependent variables, respectively. As for the independent variables, the dependent variable of model 1 (i.e., the crash pattern) was used

as a potential independent variable, but the factors that were found to be significant predictors of the crash pattern in model 1 were removed to avoid the collinearity problem. For each model, the stepwise forward selection was performed based on the Akaike information criterion (AIC) (Akaike, 1973).

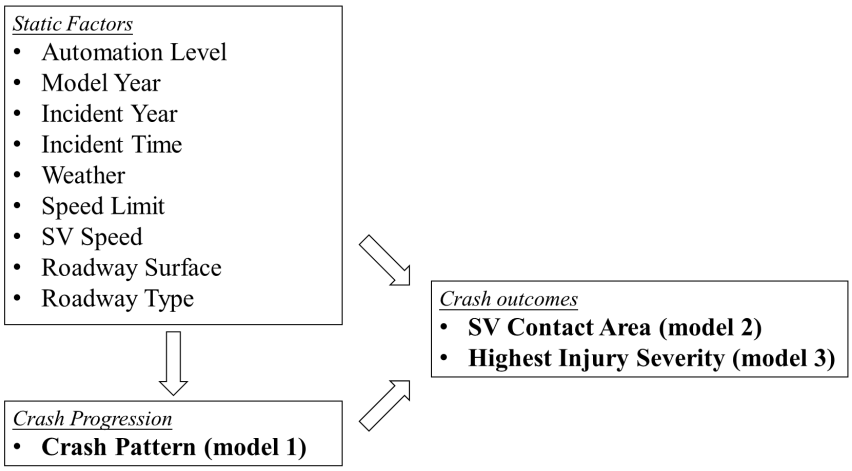

Figure 5. Hypothesized contributing factor model.

5 Results

5.1 Clustering results

The dendrogram and silhouette analysis results were used to identify the optimal number of clusters. The dendrogram in Figure 6 is a visual illustration of the clustering process. Figure 7 (a) illustrates how ASW changes with the number of clusters. The range of silhouette width is between -1 and 1, and higher silhouette width indicates better clustering performance. The number of clusters should preferably be small to ensure the interpretability of the result and avoid the overfitting problem. At the same time, we need to avoid ties in proximity (i.e., one cluster is equidistant from two or more clusters, which can be recognized in Figure 6) since it will lead to arbitrary clustering results. Based on the above considerations, the optimal number of clusters was identified as 5, with an ASW of 0.437 and most samples having a silhouette width larger than 0 (see Figure 7 (b)).

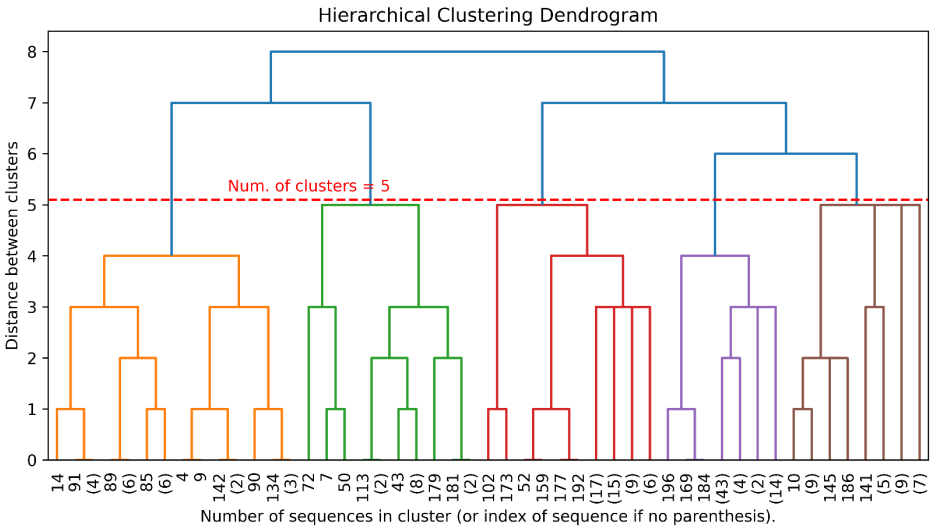

Figure 6. Dendrogram

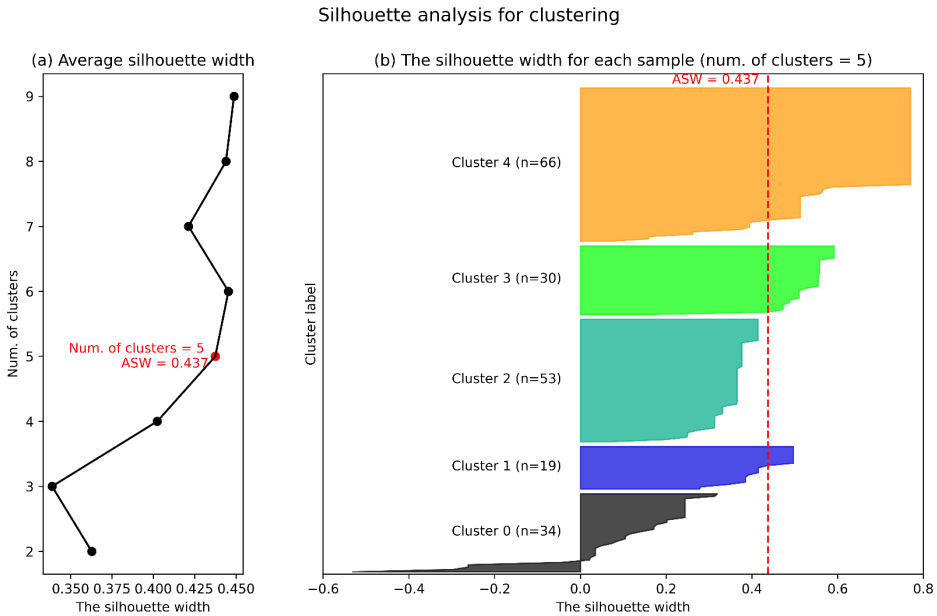

Figure 7. Silhouette analysis

Table 6. Crash patterns and the counts

Group 1 **SV maneuvers (except for stopping)**		**Group 2** **CP rule violation**		**Group 3** **CP maneuvers**	
Sequence	**Count**	**Sequence**	**Count**	**Sequence**	**Count**
D1-A2-X12	9	A1-V2-X12	8	A1-A2-X12	15
M1-A2-X12	4	A1-V2-TO-X12	3	A1-M2-X12	10
L1-A2-X12	2	R1-V2-X12	3	A1-S2-X12	8
L1-V2-X12	2	A1-TO-V2-X12	1	A1-L2-X12	4
L1-X12	2	A1-V2-R1-X13	1	A1-F-X12	3
D1-A2-X12-X13	1	A1-V2-X13	1	A1-A2-F-X12	2
D1-S1-X12	1	V2-X12	1	A1-B2-D1-X12	2
D1-TO-A2-X12	1	V2-X23-X13	1	A1-B2-X12	2
D1-X12	1	*Total*	*19 (9.4%)*	A1-F-A2-X12	2
DT-S2-X12	1			A1-E2-X12	1
DT-X12	1			A1-F-S1-X12	1
E1-X12	1			A1-M2-R2-TO-	1
F-X12	1			X12	1
L1-D1-A2-X12	1			A1-M2-TO-X12	1
L1-L2-X12	1			A1-S2-F-X12	*53 (26.2%)*
L1-S2-X12	1			*Total*	
M1-M2-X12	1				
M1-S1-A2-X12	1				
R1-D1-A2-X12	1				
V1-DT-X12	1				
Total	*34 (16.8%)*				

Group 4 **hit object**		**Group 5** **SV stopping**	
Sequence	**Count**	**Sequence**	**Count**
A1-DT-X10	7	S1-A2-X12	42
A1-X10	6	S1-B2-X12	10
A1-F-X10	5	S1-A1-A2-X12	3
R1-X10	4	A1-S1-TO-A2-X12	2
L1-X10	3	S1-V2-X12	2
L1-F-X10	2	A1-S1-A2-X12	1
A1-D1-F-X10	1	R1-S1-A2-X12	1
A1-V2-X10	1	S1-A1-X12	1
R1-F-X10	1	S1-A2-X12-X13	1
Total	*30 (14.9%)*	S1-A2-X23-X12	1
		S1-X12	1
		S1-X12-X13	1
		Total	*66 (32.7%)*

The sequences of each cluster and their counts are shown in Table 6. The pattern of each group can be identified through the common features of the sequences in the groups:

(1) Group 1: “SV maneuvres (except stop)”, in which the SV made a maneuver (D1, M1, L1, E1, R1, etc.) while the CP was proceeding straight (A2).
(2) Group 2: “CP rule violation”, in which the CP has violated the rule (V2), resulting in a crash.
(3) Group 3: “CP maneuver”, in which the CP made a maneuver (A2, M2, S2, L2, B2, E2, etc.) while the SV was going straight (A2).
(4) Group 4: “hit object”, in which the SV collided with a fixed object (X10).
(5) Group 5: “SV stopping”, in which the SV stopped (S1) and was then hit by the CP.

5.2 Frequencies of pre-crash events

To identify the difference between ADAS- and ADS-involved crashes, we conducted a comparison of pre-crash event frequencies.

Table 7 counts the frequencies of pre-crash events of ADAS and ADS across the five crash patterns.

Table 7. Frequencies of pre-crash events.

Code	ADAS		ADS		Code	ADAS		ADS	
	Count	Percentage	Count	Percentage		Count	Percentage	Count	Percentage
Group1: SV maneuvers (except for stopping)					***Group4: hit object***				
L1	6	21.4%	3	7.9%	A1	17	44.7%	3	30.0%
A2	5	17.9%	15	39.5%	F	9	23.7%	-	-
D1	3	10.7%	12	31.6%	DT	7	18.4%	-	-
DT	3	10.7%	-	-	L1	2	5.3%	3	30.0%
M1	3	10.7%	3	7.9%	D1	1	2.6%	-	-
S2	2	7.1%	-	-	R1	1	2.6%	4	40.0%
E1	1	3.6%	-	-	V2	1	2.6%	-	-
F	1	3.6%	-	-					
L2	1	3.6%	-	-					
S1	1	3.6%	1	2.6%					
V1	1	3.6%	-	-					
V2	1	3.6%	1	2.6%					
M2	-	-	1	2.6%					
R1	-	-	1	2.6%					
TO	-	-	1	2.6%					
Group2: CP rule violation					***Group5: SV stopping***				
V2	12	52.2%	7	38.9%	S1	9	56.3%	57	46.3%
A1	8	34.8%	6	33.3%	A2	5	31.3%	46	37.4%
R1	3	13.0%	1	5.6%	A1	1	6.3%	6	4.9%
TO	-	-	4	22.2%	V2	1	6.3%	1	0.8%
					B2	-	-	10	8.1%
					TO	-	-	2	1.6%
					R1	-	-	1	0.8%
Group3: CP maneuvers									
A1	34	47.2%	19	42.2%					
A2	17	23.6%	2	4.4%					
F	7	9.7%	2	4.4%					
S2	7	9.7%	2	4.4%					
M2	3	4.2%	9	20.0%					
B2	1	1.4%	3	6.7%					
E2	1	1.4%	-	-					
L2	1	1.4%	3	6.7%					
S1	1	1.4%	-	-					
D1	-	-	2	4.4%					
TO	-	-	2	4.4%					
R2	-	-	1	2.2%					

5.3 Data imputation and logistic regression models

The JS divergence of all imputed variables is given in Table 8. The range of JS divergence is between 0 and 1, with a smaller value indicating higher distributional similarity between the two samples. The result indicated that the imputed data maintained a distribution similar to that of the original data, with relatively small JS divergence values. Using the imputed data, three logistic regression models were fitted in SAS.

Table 8. Imputation performance of each variable.

Variable	JS Divergence
Incident year	<.001
Incident time	0.003
Weather	0.023
Speed limit	0.108
SV speed	0.071
Roadway surface	0.036
Roadway type	0.001
SV contact area	<.001
Highest injury severity	0.005

Error! Reference source not found. presents the type 3 Wald statistics for model 1. The automation level, speed limits, SV speed, and the interaction effect between SV speed and automation level were found to have significant effects on the crash pattern. A likelihood ratio test was conducted, and the final model was significant ($\chi^2(1) = 238.61$, $p < .0001$), with the AIC of 416.290.

Table 9. Type 3 Wald statistics analysis for model 1 (dependent variable: crash pattern)

Variable	DF	Wald Chi-Square	p-value
Automation	4	16.7724	.002
Speed Limit	4	16.6400	.002
SV Speed	4	25.2580	<.0001
SV Speed*Automation	4	10.9260	.03

Furthermore, based on the odds ratio (OR) estimates and 95% confidential interval (CI) shown in Figure 8, we found that with a higher speed limit, the pattern of a crash was less likely to be CP rule violation (OR=0.839, CI: [0.744, 0.946], p=.004) compared to the baseline pattern (i.e., hit object).

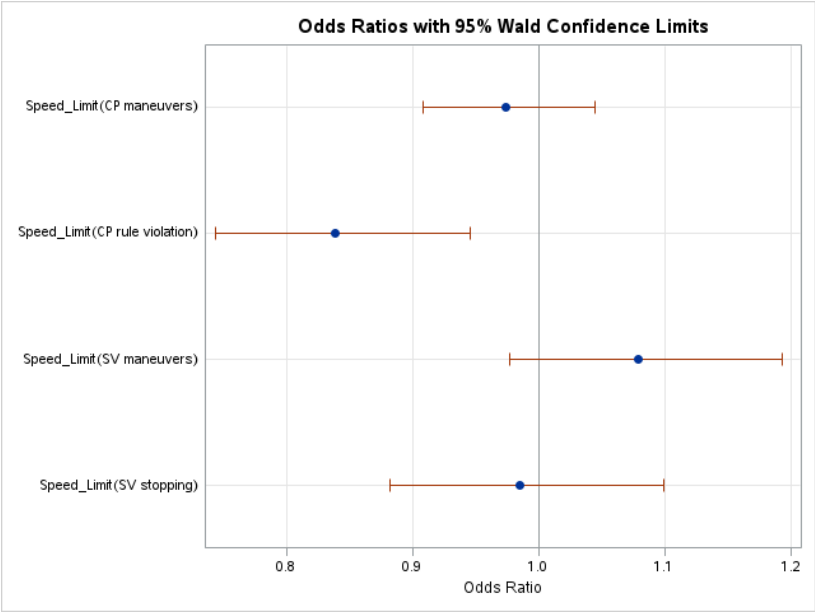

Figure 8. Odds ratio estimates for speed limit in model 1.

From Figure 9, we found that in an ADAS-involved crash, faster SV speed was associated with a lower likelihood of being SV maneuvers crash (OR=0.592, CI: [0.443, 0.792], p=.0004) and SV stopping crash (OR=0.193, CI: [0.075, 0.497], p=.0006). Similarly, in an ADS-involved crash, as SV speed increased, the crash pattern was less likely to be SV maneuvers (OR=0.869, CI: [0.775, 0.973], p=.02) or SV stopping (OR=0.676, CI: [0.566, 0.808], p<.0001), but with different odds ratio. In general, it can be observed that the association between SV speed and the crash pattern was weaker for ADS-controlled vehicles compared to that of ADAS-controlled vehicles.

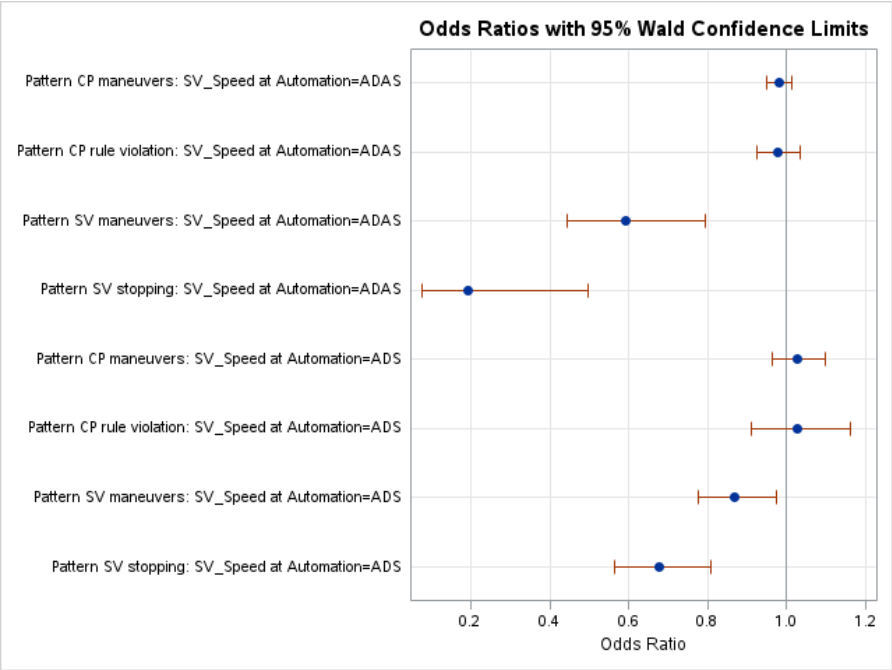

Figure 9. Odds ratio estimates for the increase of every one mph increase in SV speed given the automation level in model 1.

Error! Reference source not found. shows the type 3 Wald statistics for model 2. The results indicate that the influences of crash pattern, roadway type, model year, and incident time on SV contact were significant. A likelihood ratio test was conducted, and the final model was significant ($\chi^2(1) = 42.71$, $p < .0001$), with the AIC of 253.240.

Table 10. Type 3 Wald statistics analysis for model 2 (dependent variable: SV contact area)

Variable	DF	Wald Chi-Square	p-value
Crash Pattern	4	26.1927	<.0001
Roadway Type	1	4.9516	.03
Model Year	1	4.7347	.03
Incident Time	1	6.5731	.01

According to the odds ratio estimates in Figure 10, compared to the baseline pattern (i.e., hit object), the crash patterns of SV maneuvers (OR=12.226, CI: [3.587, 41.670], p<.0001) and SV stopping (OR=9.576, CI: [3.114, 29.448], p<.0001) were more

likely to be associated with the contact in the longitudinal direction rather than in other areas. Crashes occurring at night were more likely (OR=2.129, CI: [1.078, 4.204], *p*=.03) to involve longitudinal contact than those occurring in the daytime, and the likelihood (OR=0.384, CI: [0.185, 0.798], *p*=.01) of longitudinal contact was lower in complex roadways. Additionally, vehicles produced in 2020 or later were less likely (OR=0.447, CI: [0.220, 0.908], *p*=.03) to collide in the longitudinal direction than those produced before 2020.

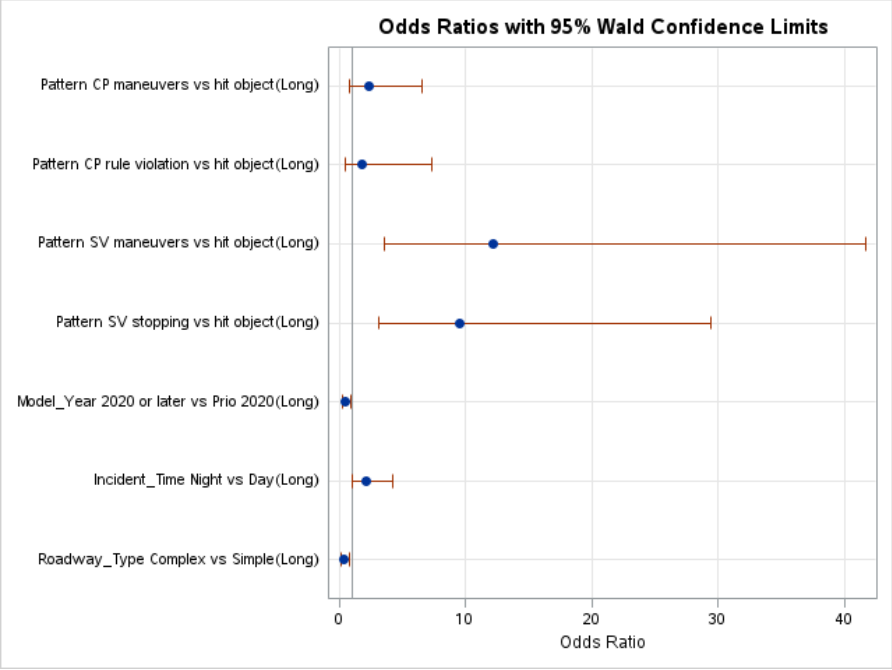

Figure 10. Odds ratio estimates for model 2 (contact area)

As for model 3, crash pattern, incident time, and roadway surface were observed to have significant effects on the highest injury severity (see Table 11). A likelihood ratio test was conducted, and the final model was significant ($\chi^2(1) = 49.10$, $p < .0001$), with the AIC of 209.068. Figure 11 demonstrates the specific impact of each variable. In terms of the impact of crash pattern, SV maneuvers (OR=7.935, CI: [2.104, 29.921], *p*=.002), CP rule violation (OR=7.917, CI: [1.971, 31.805], *p*=.004), and hit object (OR=4.494, CI: [1.432, 14.100], *p*=.01) were more likely to be associated with injuries compared to CP maneuvers. Crashes occurring at night were less likely (OR=0.338, CI: [0.142, 0.809],

p=.01) to be associated with injuries. Additionally, a wet road surface also significantly increased the probability (OR=12.575, CI: [4.329, 36.528], *p*<.0001) of having an injury.

Table 11. Wald statistics of type 3 analysis for model 3 (dependent variable: highest injury severity)

Variable	DF	Wald Chi-Square	p-value
Crash Pattern	4	17.7192	.001
Incident Time	1	5.9362	.01
Roadway Surface	1	21.6523	<.0001

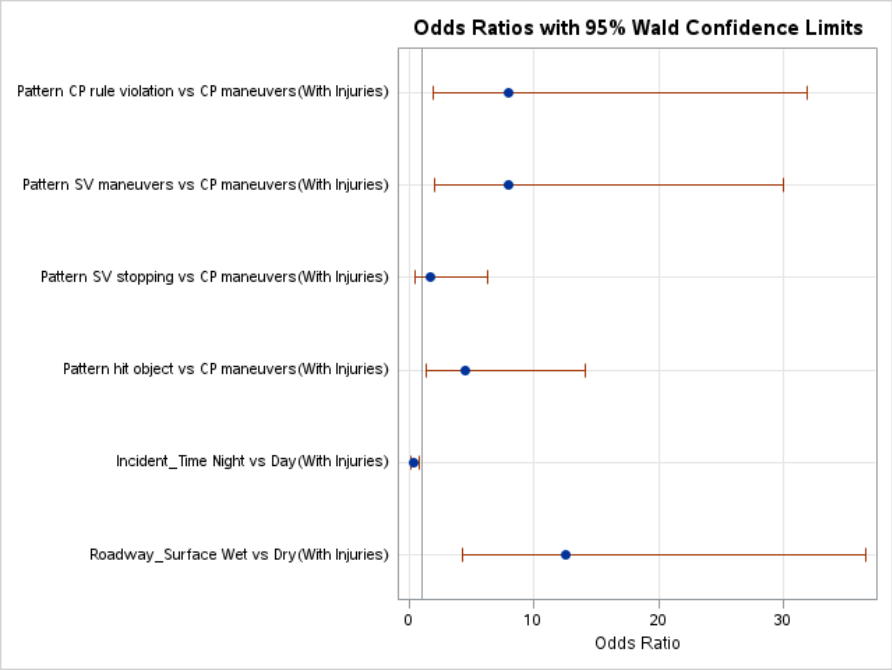

Figure 11. Odds ratio estimates for model 3 (highest injury severity)

6 Discussion

Based on the newly released NHTSA dataset, our study investigated the crash patterns of ADS- and ADAS-controlled vehicles. The sequential analysis of the events leading to a crash was conducted, and the crash patterns were categorized into five groups. Then, the factors associated with different crash patterns and the factors associated with the

outcome (i.e., contact area and the injury severity) of the crashes were analyzed.

6.1 Crash patterns and potential causes

From the clustering result of the crash sequences, this study identified five groups of crash patterns, namely SV maneuvers (except for stopping), CP rule violation, CP maneuvers, hit object, and SV stopping. The analysis of the frequency of pre-crash events in each group then allowed us to further investigate the potential causes of the crashes.

Firstly, some factors were found to be associated with both ADS- and ADAS-related crashes. For example, for SV-maneuvers-related-crashes (i.e., Group 1), SV decelerates (D1) was common for both ADAS and ADS, which may explain the high likelihood of ADS/ADAS-controlled vehicle being rear-ended (Huang, Wen, & He, 2022).

However, our research also highlights the necessity to differentiate the ADS and ADAS crashes. For example, the lateral maneuvers of the SV, including left turning (L1) and merging (M1), were found to be frequent in ADAS-involved crashes but not in ADS-involved crashes. Usually, ADAS had weaker perception and motion planning capability compared to ADS. Given that the drivers usually have a weak mental model of driving automation that handles lateral motion control (e.g., lane keeping assist) (Naujoks et al., 2017; Huang et al., 2023), drivers may not be able to correctly identify the situations that they should be responsible for. These two factors may have explained the high frequency of lateral maneuver (i.e., L1 and M1) among ADAS-controlled vehicles in group 1.

At the same time, what is alerting is that we found that driver distraction (D1) was common in ADAS crashes, especially in crashes related to SV maneuvers and in hit-object crashes. This finding indicates that the engagement of distraction tasks may be prevalent in ADAS-controlled vehicles, at least when a crash happens. This finding echoes the results in previous simulator studies, which found an increased likelihood of

distraction engagement with ADAS compared to that in non-automated vehicles (He & Donmez, 2019), but our study provides additional information on the potential outcome of distracted driving in ADAS-controlled vehicles. Thus, interventions to prevent distracted driving are important for improving the safety performance of ADAS.

To further reveal the factors leading to different types of crashes, a logistic regression model was built to identify the association between potential contributing factors and crash patterns. The contributing factor model of the crash pattern (model 1) suggested that the automation level, speed limit, and SV speed are associated with the crash patterns. The result also showed that for both ADS- and ADAS-controlled vehicles, the crashes related to SV maneuvers or SV stopping tended to occur at lower SV speeds, whereas crashes involving hitting objects tended to occur at higher SV speeds. Such a trend was more obvious among ADAS-controlled vehicles. One potential explanation is that, for ADS-controlled vehicles, as has been found by Huang et al. (2022), drivers tended to keep a smaller headway distance when following ADS-controlled vehicles than that when following human-driven vehicles at low speeds. At the same time, as also been found by Huang et al. (2022), the ADS may still not be capable of anticipating the hazards and thus may take maneuvers unexpectedly (as suggested by smaller time-to-collision when drivers were following ADS-controlled vehicles), which increases the chance of ADS involving in a crash. The ADAS-controlled vehicle, however, may even be less capable of being proactive or responsive than ADS-controlled vehicles, especially at lower speeds when the traffic situation is complex, and thus, their behavior is even less predictable than that of ADS-controlled vehicles. Although the speed-crash relationship has been investigated in many existing studies, a common perspective focuses on the effect of speed on crash severity (Imprialou et al., 2016). The results of our study have

therefore provided a new perspective on this relationship by identifying a correlation between SV speed and crash pattern.

6.2 Contributing factors of crash outcomes

Based on two additional logistic regression models, we have identified how the crash pattern, in combination with other crash contributing factors, is associated with the crash outcome (i.e., contact area and injury severity). In general, we found that out of all crashes with longitudinal contact, 70% of them involved ADS- or ADAS-controlled vehicles being rear-ended by a following vehicle (i.e., with the contact area of the rear on SV). Previous research has pointed out that the ADS- or ADAS-controlled vehicles were highly likely to be rear-ended by the following vehicles (Huang, Wen, & He, 2022; Liu et al., 2021) and pointed out that the motion-related factors (e.g., vehicle speed) were highly related with the odds of being rear-ended compared to that of environment-related factors (e.g., weather) (Huang, Wen, & He, 2022). Using a logistic model based on crash sequence analysis (model 2), our study further revealed the temporal process before potential rear-ending crashes. Specifically, the crash patterns of SV maneuvers and SV stopping were more likely to be associated with a crash contact in the longitudinal direction.

In addition, we found that a number of environmental factors may moderate the likelihood of rear-end crashes. For example, crashes occurring at night were more likely to involve longitudinal contact than those occurring in the daytime, potentially because the dark environment can impair the perception capability of the SVs, leading to more unexpected sudden maneuvers of the SVs and thus increased the chance of being rear-ended by other vehicles. Further, vehicles were more likely to be involved in crashes with longitudinal contact on simple roadways compared to that on complex roadways, likely because of the lower need for lateral motions on simple roadways. The optimization of

the ADS/ADAS control algorithms or the training of the human drivers (e.g., to keep a larger distance with ADS- or ADAS-controlled vehicles) in mixed traffic may help reduce the crashes with ADS- and ADAS-controlled vehicles. For example, we found that the car models produced in 2020 or later are less likely to be involved in a longitudinal collision, which is possibly due to the improvement in longitudinal collision intervention features (e.g., front collision warning, automatic emergency braking) in recent years (Highway Loss Data Institute, 2022).

At the same time, the crash pattern, incident time, and roadway surface were found to be associated with the highest injury severity (model 3). It was found that certain patterns of crashes (i.e., SV maneuvers, CP rule violation, and hit object) were more likely to result in injuries. Specifically, the CP rule violation was most likely to cause an injury. This is potential because the SV may not be able to respond well to CP that did not obey the right-of-way on the road, and this may have led to collisions at relatively high speeds. Furthermore, the results indicate that crashes occurring at night were less likely to result in injuries compared to the ones that happened in the daytime, which contradicts the conclusions from previous studies based on the CA DMV dataset, in which autonomous vehicle crashes were found to be more fatal at night (Kutela et al., 2022). This difference might be attributed to the development of vehicle perception technologies in recent years, but further investigation is needed. Another finding is that a wet road surface significantly increased the probability of injury. This is as expected and consistent with previous research that has shown that wet road conditions can contribute to crashes and injuries (Kutela et al., 2022).

6.3 Limitations

Due to the limited number of available crash reports, the sample size of this study was unbalanced and relatively small. To make full use of the data, the missing values in the

reports were imputed using the MissForest algorithm. Although the reliability of this imputation algorithm has been verified by existing studies, and the imputed data was found to follow similar distributions as the original data, it should be noted that the imputed data may not completely reflect the actual characteristics of the crashes. Given that the functionality of automated driving is evolving dramatically, when more ADS- and ADAS-involved crash data becomes available, updated analyses should be conducted based on datasets with better quality to keep track of the latest trends in ADS- and ADAS-involved crashes.

7 Conclusions

In conclusion, this study has identified five distinct crash patterns and analyzed their potential causes. By utilizing logistic regression models, the study determined that automation level, speed limit, and SV speed are associated with crash patterns. Additionally, the study found that crash patterns, together with other contributing factors such as roadway type, model year, incident time, and roadway surface, may lead to different crash outcomes (i.e., SV contact areas and the highest injury severity). In general, our finding suggests that the ADS and ADAS systems are still less able to handle some on-road incidents well (e.g., the rule violation of other road agents) compared to human drivers. Further, the development of driving automation technology has led to changes in crash characteristics (as different conclusions were drawn from CA DMV and NHTSA datasets). Lastly, the event sequence analysis can provide more information regarding the causes of crashes compared to a pure comparison of static information (e.g., weather and speed limit). Driver training and the optimization of the algorithms are still needed to improve the safety of mixed traffic with ADS-controlled, ADAS-controlled, and human-driven vehicles.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 52202425), Guangzhou Municipal Science and Technology Project (Project No. 2023A03J0011), Guangzhou Science and Technology Program City-University Joint Funding Project (Project No. 2023A03J0001) and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083). The authors would express their most sincere thanks to the National Highway Traffic Safety Administration (NHTSA) for their efforts in collecting and releasing the raw data used in this study.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle–In: Second International Symposium on Information Theory (Eds) BN Petrov, F. Csaki. BNPBF Csaki Budapest: Academiai Kiado.

Arribas-Gil, A., & Müller, H.-G. (2014). Pairwise dynamic time warping for event data. Computational Statistics & Data Analysis, 69, 255–268.

California DMV. (2022). *Autonomous Vehicle Collision Reports*. California Department of Motor Vehicles (California DMV). https://www.dmv.ca.gov/portal/vehicle-industry-services/ autonomous-vehicles/autonomous-vehicle-collision-reports/

Chen, H., Chen, H., Liu, Z., Sun, X., & Zhou, R. (2020). Analysis of factors affecting the severity of automated vehicle crashes using XGBoost model combining POI data. *Journal of Advanced Transportation*, *2020*, e8881545.

Cheng, S., Li, L., Guo, H.-Q., Chen, Z.-G., & Song, P. (2020). Longitudinal collision avoidance and lateral stability adaptive control system based on MPC of autonomous

vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *21*(6), 2376–2385.

Consumer Reports. (2021). *Understanding the Current State of Vehicle Automation*. Consumer Reports. https://data.consumerreports.org/insights-and-impact/

Dadvar, S., & Ahmed, M. M. (2021). California autonomous vehicle crashes: Explanatory data analysis and classification tree (TRBAM-21-04068). *Transportation Research Board 100th Annual Meeting*, Washington DC, United States.

Das, S., Dutta, A., & Tsapakis, I. (2020). Automated vehicle collisions in California: Applying Bayesian latent class model. *IATSS Research*, *44*(4), 300–308.

Ding, S., Abdel-Aty, M., Wang, D., Barbour, N., Wang, Z., & Zheng, O. (2023). Exploratory analysis of injury severity under different levels of driving automation (SAE Level 2-5) using multi-source data (arXiv:2303.17788). arXiv. https://doi.org/10.48550/arXiv.2303.17788

Esenturk, E., Khastgir, S., Wallace, A., & Jennings, P. (2021). Analyzing real-world accidents for test scenario generation for automated vehicles. 2021 IEEE Intelligent Vehicles Symposium (IV), 288–295.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 96(34), 226–231.

Galloway, A. J., Bareiss, M., Hasegawa, T., Sherony, R., & Riexinger, L. E. (2023). Evaluation of intersection crashes using naturalistic driving data through the lens of future I-ADAS. Traffic Injury Prevention, 24(7), 577–582.

Goodall, N. (2023). Comparability of Automated Vehicle Crash Databases (arXiv:2308.00645). arXiv. https://doi.org/10.48550/arXiv.2308.00645

He, D., & Donmez, B. (2019). Influence of driving experience on distraction engagement in automated vehicles. *Transportation Research Record*, *2673*(9), 142–151.

Highway Loss Data Institute. (2022). Predicted availability of safety features on registered vehicles - A 2022 update. *Bulletin*, *2*(39).

Huang, C., Wen, X., & He, D. (2022). Characteristics of rear-end collisions: A comparison between ADS-involved crashes and ADAS-involved crashes (TRBAM-23-00506). *Transportation Research Board 102th Annual Meeting*, Washington DC, United States.

Huang, C., Wen, X., He, D., & Jian, S. (2022). Sharing the road: How human drivers interact with autonomous vehicles on highways. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1437–1441.

Huang, C., He, D., Wen, X., & Yan, S. (2023). Beyond adaptive cruise control and lane centering control: Drivers' mental model of and trust in emerging ADAS technologies. *Frontiers in Psychology*, 14.

Imprialou, M.-I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash–speed relationships: A new perspective in crash modelling. *Accident Analysis & Prevention*, *86*, 173–185.

Kaber, D., Zhang, Y., Jin, S., Mosaly, P., & Garner, M. (2012). Effects of hazard exposure and roadway complexity on young and older driver situation awareness and

performance. *Transportation Research Part F: Traffic Psychology and Behaviour*, 15(5), 600–611.

Kutela, B., Avelar, R. E., & Bansal, P. (2022). Modeling automated vehicle crashes with a focus on vehicle at-fault, collision type, and injury outcome. *Journal of Transportation Engineering, Part A: Systems*, *148*(6), 04022024.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Liu, Q., Wang, X., Wu, X., Glaser, Y., & He, L. (2021). Crash comparison of autonomous and conventional vehicles using pre-crash scenario typology. *Accident Analysis & Prevention*, *159*, 106281.

Lu, J. J., Chen, S., Ge, X., & Pan, F. (2013). A programmable calculation procedure for number of traffic conflict points at highway intersections. *Journal of Advanced Transportation*, 47(8), 692–703.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281–297.

Müllner, D. (2011). *Modern Hierarchical, Agglomerative Clustering Algorithms* (arXiv:1109.2378). arXiv. http://arxiv.org/abs/1109.2378

Naujoks, F., Purucker, C., Wiedemann, K., Neukum, A., Wolter, S., & Steiger, R. (2017). Driving performance at lateral system limits during partially automated driving. *Accident Analysis & Prevention*, *108*, 147–162.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search

for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

National Highway Traffic Safety Administration. (2022). *Standing General Order on Crash Reporting*. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting.

NHTSA. (2023). Crash Data Systems. National Highway Traffic Safety Administration (NHTSA). https://www.nhtsa.gov/data/crash-data-systems

Ren, W., Yu, B., Chen, Y., & Gao, K. (2022). Divergent effects of factors on crash severity under autonomous and conventional driving modes using a hierarchical Bayesian approach. *International Journal of Environmental Research and Public Health*, *19*(18), 11358.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

SAE International. (2021). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE J3016).

Scanlon, J. M., Kusano, K. D., Sherony, R., & Gabler, H. C. (2015). Potential of intersection driver assistance systems to mitigate straight crossing path crashes using U.S. nationally representative crash data. *2015 IEEE Intelligent Vehicles Symposium (IV)*, 1207–1212.

Scanlon, J. M., Sherony, R., & Gabler, H. C. (2016). Preliminary potential crash prevention estimates for an Intersection Advanced Driver Assistance System in straight crossing path crashes. *2016 IEEE Intelligent Vehicles Symposium (IV)*, 1135–

1140.

Seacrist, T., Maheshwari, J., Sarfare, S., Chingas, G., Thirkill, M., & Loeb, H. S. (2021). In-depth analysis of crash contributing factors and potential ADAS interventions among at-risk drivers using the SHRP 2 naturalistic driving study. *Traffic Injury Prevention*, 22(sup1), S68–S73.

Song, Y., Chitturi, M. V., & Noyce, D. A. (2021). Automated vehicle crash sequences: Patterns and potential uses in safety testing. *Accident Analysis & Prevention*, *153*, 106017.

Song, Y., Chitturi, M. V., & Noyce, D. A. (2022). Intersection two-vehicle crash scenario specification for automated vehicle safety evaluation using sequence analysis and Bayesian networks. *Accident Analysis & Prevention*, *176*, 106814.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Torres, J., Li, Y., & Zhang, J. (2021). Investigating traffic crashes involving autonomous vehicles (10311046). *Proceedings of the 2021 IISE Annual Conference*, 1046-1051.

Wen, X., Jian, S., & He, D. (2023). Modeling the Effects of Autonomous Vehicles on Human Driver Car-Following Behaviors Using Inverse Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 1–13.

Wu, K.-F., Thor, C. P., & Ardiansyah, M. N. (2016). Identify sequence of events likely to result in severe crash outcomes. *Accident Analysis & Prevention*, *96*, 198–207.

Xu, C., Ding, Z., Wang, C., & Li, Z. (2019). Statistical analysis of the patterns and

characteristics of connected and autonomous vehicle involved crashes. *Journal of Safety Research*, *71*, 41–47.

Zhang, Z., Wang, C., Zhao, W., & Feng, J. (2022). Longitudinal and lateral collision avoidance control strategy for intelligent vehicles. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *236*(2–3), 268–286.

Zheng, O., Abdel-Aty, M., Wang, Z., Ding, S., Wang, D., & Huang, Y. (2023). AVOID: Autonomous Vehicle Operation Incident Dataset Across the Globe (arXiv:2303.12889). *arXiv*. https://doi.org/10.48550/arXiv.2303.12889