US DOT National
University Transportation Center for Safety

Analysis of Contributing Factors in Crashes Involving Electric Vehicles and Vehicles with Warning Systems and Level 1 Automated Features: A State Level Analysis

Corey Harper (PI), https://orcid.org/0000-0003-1956-5258

Chris Hendrickson (Co-PI), https://orcid.org/0000-0002-9812-3580

Jiacheng Wang

Final Report – 08/31/2026

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under [grant number 69A3552344811 / 69A3552348316] from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Analysis of Contributing Factors in Crashes Involving Electric Vehicles and Vehicles with Warning Systems and Level 1 Automated Features: A State Level Analysis		**5. Report Date** July 31, 2026 **6. Performing Organization Code**
7. Author(s) Corey Harper (PI), https://orcid.org/0000-0003-1956-5258 Chris Hendrickson (Co-PI), https://orcid.org/0000-0002-9812-3580 Jiacheng Wang		**8. Performing Organization Report No.** **584**

9. Performing Organization Name and Address Carnegie Mellon University 5000 Forbes Ave. Pittsburgh PA 15213	**10. Work Unit No.**
	11. Contract or Grant No. No. 69A3552344811 / No, 69A3552348316 Safety 21
12. Sponsoring Agency Name and Address Safety 21 University Transportation Center Carnegie Mellon University 5000 Forbes Ave. Pittsburgh PA 15213	**13. Type of Report and Period Covered** Final Report (July 1, 2025 – July 31, 2026)
	14. Sponsoring Agency Code
15. Supplementary Notes Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.	

16. Abstract

Automated driving features could improve road safety by reducing the number of crashes caused by human error. However, as the market penetration of Level 2 advanced driver assistance systems (ADAS) increases, so does the number of crashes involving these technologies. This study proposes a data-driven framework combining latent class analysis, association rule mining, and logistic regression to analyze 617 Level 2 ADAS crashes reported to the National Highway Traffic Safety Administration between 2019 and 2026. Two latent classes are identified, which are lower-speed local road crashes and high-speed highway crashes. Association rule mining and logistic regression are conducted within each latent class, revealing substantial heterogeneity in injury-related crash patterns. In lower-speed local road crashes, injury occurrences are mainly associated with frontal impacts, intersection-related crashes, and moderate-speed configurations. In high-speed highway crashes, injury occurrences are mainly associated with high-speed configurations, fixed objects, light-to-medium duty vehicles, frontal impacts, and morning or late-night conditions. Practical applications include placing greater emphasis in scenario-based testing for combinations involving moderate and high pre-crash speeds, frontal contact, fixed-object crashes, light-to-medium duty vehicle crash partners, and late-night conditions. These findings may help manufacturers refine speed-aware system-use guidance, driver warnings, and operational design constraints for Level 2 ADAS-equipped vehicles. Overall, this study demonstrates the value of class-specific analysis for uncovering heterogeneous injury-associated patterns in Level 2 ADAS crashes.

17. Key Words	**18. Distribution Statement**
Level 2 ADAS; Latent Class Analysis; Association Rule Mining; Heterogeneity	

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	25	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Problem Statement

Based on the society of Automotive Engineers (SAE), driving automation is classified into six levels (SAE, 2021). SAE Levels 1–2 are commonly characterized as Advanced Driver Assistance Systems (ADAS), whereas SAE Levels 3–5 are generally referred to as Automated Driving Systems (ADS). SAE Level 2 ADAS has been deployed in consumer vehicles, with manufacturers such as Tesla and GM offering these systems in commercially available models. These systems are expected to improve the performance and safety of transportation systems, by lowering the number of crashes caused by human error (Khan et al.,2019; Harper et al., 2016). However, as the market penetration of Level 2 ADAS increases, so does the absolute number of crashes involving these technologies. To ensure that the anticipated safety benefits of automation are realized in practice, it is crucial for policymakers, infrastructure owners and operators, and technology providers to understand the factors contributing to crashes involving these technologies. This understanding is necessary to assess existing safety risks and identify areas where additional testing, algorithm development, best-practice sharing, and infrastructure improvements are needed. In 2021, the National Highway Traffic Safety Administration (NHTSA) issued a Standing General Order requiring manufacturers and operators to report crashes involving vehicles equipped with ADS or Level 2 ADAS (NHTSA, 2021). This requirement ensures that NHTSA receives timely and standardized information on safety-critical incidents, enabling systematic investigation, regulatory oversight, and enforcement actions related to automated driving technologies.

This study develops a hybrid, data-driven analytical framework that integrates latent class analysis (LCA), association rule mining (ARM) and logistic regression to better understand contributing factors in crashes involving vehicles equipped with Level 2 ADAS. LCA is employed to identify homogeneous latent classes of ADAS-involved crashes based on multidimensional crash characteristics, while ARM is subsequently applied within each latent class to extract representative crash scenarios and co-occurring factors associated with injury outcomes. Logistic regression is further used as a supervised validation and explanatory modeling tool to quantify the statistical associations between injury outcome and crash characteristics. Using this framework, the study addresses three research questions. First, what distinct latent classes of ADAS-involved crashes can be identified based on diverse crash characteristics? Second, within each latent class, which combinations of crash-related factors are most strongly associated with injury outcomes, as revealed by association rules with high lift and thorough logistic regression? Third, how do injury-related crash scenarios differ across latent class, and what do these differences indicate about heterogeneous risk patterns underlying ADAS-related crashes? Overall, by answering these questions, this study develops a data-driven framework to identify heterogeneous crash scenarios and reveal their risk patterns, thereby informing areas for future testing and technology development to improve Level 2 ADAS safety and performance.

Data and Method

Data Preparation

This study utilizes standing general order summary incident report data reported to NHTSA under the standing general order (NHTSA, 2021). This analysis uses the data files containing incidents involving ADAS equipped vehicles. To ensure data integrity and comparability, the same data cleaning and preprocessing procedures outlined in Wang et al. (2026) were applied. These include removing duplicate report versions submitted by the same reporting entity by retaining only the most recent report version, eliminating duplicate records arising from multiple entities reporting the same incident, and consolidating records so that each unique vehicle involved in an incident is represented only once. Feature engineering procedures, including the recategorization of variables with numerous unique levels (e.g., pre-crash maneuvers, crash partner types, and speed-related attributes), were also conducted following the same methodology. Injury outcome, as indicated by the "Highest Injury Severity Alleged" field, is used as the dependent outcome variable in this study. By adopting an identical data processing framework, this study ensures methodological consistency with prior work while enabling a focused investigation of crash patterns and injury-related scenarios specific to ADAS-equipped vehicles. Only crashes with available injury outcome information were retained for further analysis.

After data preprocessing, 617 Level 2 ADAS crash records with available injury outcome information were retained for analysis. The variables used in this study include speed (MPH), weather condition, time of day, roadway type, roadway surface, subject vehicle pre-crash movement, crash partner, subject vehicle contact area, subject vehicle mileage driven, and subject vehicle model year. Here, the subject vehicle refers to the ADAS-equipped vehicle involved in the crash. Table 1 presents the count and percentage distribution of these variables across three injury severity levels: no injuries reported, minor/moderate injuries reported, and serious injuries/fatality reported. The dataset includes 312 crashes with no injuries reported, 185 crashes with minor or moderate injuries reported, and 120 crashes involving serious injuries or fatalities. The incident dates of the analyzed crashes span from August 2019 to March 2026.

Based on Table 1, ADAS-involved crashes in the dataset were most frequently observed at higher pre-crash speeds, particularly above 60 mph. Highway/freeway crashes accounted for the largest share across all injury severity levels, representing 54.2% of no-injury crashes, 55.1% of minor/moderate injury crashes, and 57.5% of serious injury/fatal crashes. Clear weather and dry roadway surface conditions were also common, suggesting that many ADAS-involved crashes occurred under generally normal environmental conditions rather than adverse weather or roadway surface conditions. Moving straight or passing was the predominant pre-crash movement of the subject vehicle, especially among minor/moderate injury crashes. The front of the subject vehicle was the most common contact area, accounting for 60.0% of minor/moderate injury crashes and 41.7% of serious injury/fatal crashes. Crash partners varied by injury severity: light-to-medium duty vehicles and heavy-duty vehicles accounted for larger shares of injury crashes than no-injury crashes, while non-motorist and motorcycle crashes were particularly overrepresented among serious injury/fatal crashes. Most subject vehicles were manufactured in or after 2020, although vehicles manufactured before 2020 accounted for a higher proportion of serious injury/fatal crashes than no-injury crashes.

Table 1. Characteristics of ADAS-Involved Crashes Used For the Analysis

Features		**No Injuries Reported Count (%)**	**Minor/Moderate Injuries Reported Count (%)**	**Serious Injuries/Fatality Reported Count (%)**
Speed (MPH)	0-20	51 (16.4%)	18 (9.73%)	13 (10.8%)
	20-40	46 (14.7%)	21 (11.3%)	8 (6.67%)
	40-60	50 (16%)	34 (18.4%)	23 (19.2%)
	>60	90 (28.9%)	70 (37.8%)	40 (33.3%)
Weather Condition	Clear	137 (43.9%)	92 (49.7%)	52 (43.3%)
	Cloudy	41 (13.1%)	17 (9.19%)	8 (6.67%)
	Adverse	16 (5.13%)	16 (8.65%)	6 (5%)
Time of Day	Morning	55 (17.6%)	41 (22.2%)	32 (26.7%)
	Afternoon	80 (25.6%)	51 (27.6%)	15 (12.5%)
	Evening	67 (21.5%)	28 (15.1%)	30 (25%)
	Late Night	51 (16.4%)	26 (14.1%)	24 (20%)
Roadway Type	Highway / Freeway	169 (54.2%)	102 (55.1%)	69 (57.5%)
	Street	47 (15.1%)	26 (14.1%)	12 (10%)
	Intersection	21 (6.73%)	26 (14.1%)	9 (7.5%)
	Parking Lot	8 (2.56%)	3 (1.62%)	1 (0.83%)
	Rural Road	4 (1.28%)	9 (4.86%)	2 (1.67%)
Roadway Surface	Dry	151 (48.4%)	97 (52.4%)	59 (49.2%)

	Wet / Snow / Slush / Ice	18 (5.77%)	20 (10.8%)	8 (6.67%)
Subject Vehicle Pre-Crash Movement	Moving Straight or Passing	168 (53.9%)	119 (64.3%)	56 (46.7%)
	Lane Changes and Merging	4 (1.28%)	7 (3.78%)	4 (3.33%)
	Turning	17 (5.45%)	12 (6.49%)	1 (0.83%)
	Stopped/Parked/ Backing	13 (4.17%)	6 (3.24%)	5 (4.17%)
Crash Partner	Light-to-Medium Duty Vehicle	57 (18.3%)	50 (27%)	36 (30%)
	Heavy Duty Vehicle	12 (3.85%)	14 (7.57%)	11 (9.17%)
	Non-Motorist & Motorcycle	7 (2.24%)	3 (1.62%)	16 (13.3%)
	Fixed Object	115 (36.9%)	62 (33.5%)	25 (20.8%)
	Animal	23 (7.37%)	7 (3.78%)	0 (0%)
Subject Vehicle Contact Area	Front	119 (38.1%)	111 (60%)	50 (41.7%)
	Side	53 (17%)	24 (13%)	5 (4.17%)
	Rear	33 (10.6%)	18 (9.73%)	15 (12.5%)
	Other	28 (8.97%)	3 (1.62%)	1 (0.83%)
Subject Vehicle Mileage Driven	0-25,000	130 (41.7%)	45 (24.3%)	39 (32.5%)
	25,000-50,000	50 (16%)	27 (14.6%)	13 (10.8%)
	>50,000	26 (8.33%)	21 (11.3%)	16 (13.3%)
Subject Vehicle Model Year	Before 2020	36 (11.5%)	23 (12.4%)	26 (21.7%)
	In/After 2020	276 (88.5%)	161 (87%)	94 (78.3%)

Latent Class Analysis

LCA is a statistical method for discovering separate subgroups of a dataset based on mutual characteristics and response patterns. In this study, these subgroups are referred to as latent classes. LCA has been used extensively in traffic safety research to address heterogeneity in crash data and to identify homogeneous crash groups before conducting further analysis (e.g., Wang et al., 2026). Suppose we have I crashes and J indicators. Every crash is considered as a unit. The probability of a particular response pattern for crash i can be expressed as:

$$P(X_i = x_i) = \sum_{n=1}^{N} P(C_n) \prod_{j \in O_i} P(X_{ij} = x_{ij} \mid C_n) \quad (1)$$

Where, $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$ is the set of variables for unit i, x_i = a specific response pattern observed for unit i, N = the number of latent classes, C_n = the n-th latent class, $P(C_n)$ = membership probability of latent class C_n, $P(X_{ij} = x_{ij} \mid C_n)$ = the conditional probability of unit i exhibiting response x_{ij} to indicator j, given latent class C_n. These J indicators include the variables reported in Table 1.

In this study, LCA was implemented using the Python StepMix package (Morin et al., 2025). We used the 'categorical_nan' measurement model, which handles missing values directly through full-information maximum likelihood (FIML). Under this approach, the probability of a particular response pattern for crash i is computed using only the indicators observed for that crash. Therefore, we define $O_i \subseteq \{1, \ldots, J\}$ as the set of observed indicators for crash i, and missing indicators are excluded from the likelihood product rather than imputed. In this framework, LCA estimates the model parameters that maximize the likelihood of the observed data using the Expectation–Maximization algorithm. After fitting the LCA model, each crash is assigned to a latent class using Bayes' theorem to calculate the posterior probabilities of belonging to each latent class (eq. 2):

$$P(C_n \mid X_i = x_i) = \frac{P(C_n) \prod_{j \in O_i} P(X_{ij} = x_{ij} \mid C_n)}{P(X_i = x_i)} \quad (2)$$

where, $P(C_n \mid X_i = x_i)$ = the probability that unit i is in latent class C_n, given its observed response pattern x_i.

The performance of the fitted model is assessed using the information criteria such as the AIC (Akaike, 1974), BIC (Schwarz, 1978), and Consistent Akaike Information Criterion (CAIC) (Weller et al, 2020). AIC is a commonly used measure for model selection. AIC balances model fit and complexity by penalizing the likelihood based on the number of free parameters in the model. Lower AIC values indicates a better balance between goodness-of-fit and simplicity. BIC, similarly to AIC, evaluates the trade-off between model fit and complexity, but it imposes a penalty that increases with sample size (Padgett and Tipton, 2020). The CAIC, further extends AIC to account for larger sample sizes and adds an additional penalty based on the logarithm of the number of observations. Together, these metrics guide the selection of the model by indicating which model is most likely to be correct relative to others, based on its structure and the data it seeks to explain. Lower values across these criteria generally suggest a preferred model, indicating a balance between accuracy and simplicity.

Association Rule Mining

Association Rule Mining is a data mining technique used to identify meaningful associations and co-occurrence patterns among variables in large datasets and has been widely applied in transportation crash studies (Labbo et al.,2024; Liu et al.,2024; Fagerlind et al., 2022). The association rule is defined as follows: Suppose $I = \{I_1, I_2, \ldots\ldots I_m\}$ represents the itemset, which contains all unique binary items generated from the crash characteristic variables listed in Table 1. Each item is a unique value of a variable. For instance, the Time of Day variable can be converted into 4 binary items which are Morning, Afternoon, Evening, and Late Night, and a similar encoding procedure is applied to all other variables. Every ADAS crash record corresponds to a transaction that contains a subset of I. An association rule has the form of $X \Longrightarrow Y$, where X is the antecedent and Y is the consequent and both X and Y are a subset of I. ARM evaluates the strength of rules using three primary metrics which are support, confidence and lift. Their mathematical definitions are shown in Equations 3-5. Support measures the proportion of transactions containing a given itemset. Confidence represents the conditional probability of observing Y given X. Lift quantifies the strength of association between X and Y relative to their independent occurrence.

$$\text{Support (X)} = \text{P(X)} = \frac{Count\ (X)}{N} \qquad (3.1)$$

$$\text{Support (Y)} = \text{P(Y)} = \frac{Count\ (Y)}{N} \qquad (3.2)$$

$$\text{Support } (X \Longrightarrow Y) = P(X \cap Y) = \frac{Count\ (X \cap Y)}{N} \qquad (3.3)$$

$$\text{Confidence } (X \Longrightarrow Y) = P(Y \mid X) = \frac{Support\ (X \Rightarrow Y)}{Support\ (X)} \qquad (4)$$

$$\text{Lift } (X \Longrightarrow Y) = \frac{\text{Confidence } (X \Rightarrow Y)}{Support\ (Y)} \qquad (5)$$

In the equations above, Count(X) is the number of crashes with X; Count(Y) is the number of crashes with Y; N = total number of crashes; count $(X \cap Y)$ is the number of crashes with both x and y.

For this study, the Apriori algorithm was employed to mine frequent itemsets and subsequently generate association rules under predefined minimum support and confidence thresholds. Injury severity was first binarized into injury versus no injury. Association rules were then mined within each latent class, and only rules with injury as the consequent were retained for interpretation. This design enables the identification of contributing factors and crash scenarios associated with injury outcomes within each latent class. A number of prior transportation crash studies using association rule mining have relied on a trial-and-error approach to determine the minimum support and confidence thresholds (e.g., Labbo et al., 2024; Hossain et al., 2022; Liu et al., 2024; Tamakloe et al., 2025). To improve the robustness of the discovered rules, this study employed a two-stage robustness analysis that combines threshold sensitivity analysis with bootstrap resampling. First, for each minimum support-confidence threshold combination, we generated 500 bootstrapped samples by random sampling with replacement from the class-specific dataset. ARM was applied to each bootstrap sample using the same support and confidence thresholds. Within each threshold combination, only rules that reappeared in at least 80% of the bootstrap samples were retained as bootstrap-stable rules. This step was used to remove rules that were highly sensitive to sampling variation and this procedure was adapted from the bootstrapped ARM workflow described by Jamison (2024) with modifications to the rule-retention criterion. Second, we conducted a threshold sensitivity analysis across multiple support-confidence combinations. Specifically, we evaluated nine threshold combinations formed by three minimum support values and three minimum confidence values. After applying the bootstrap stability criterion within each threshold combination, we summarized how many times each bootstrap-stable rule appeared across the nine threshold settings. Rules were prioritized for interpretation only if they appeared in at least six of the nine threshold combinations, corresponding to an across-threshold existence rate of at least 66.7%.

Logistic Regression

To further assess the statistical relevance of the factors and factor combinations identified in the preceding analyses, binary logistic regression models were estimated with injury occurrence as the dependent variable. Injury crashes were coded as 1, and non-injury crashes were coded as 0. The probability of an injury crash for observation i, $P(Y_i = 1)$, was modeled as follows:

$$P(Y_i=1) = \frac{1}{1 + e^{-\eta_i}} \quad (6)$$

$$\eta_i = \beta_0 + X_{1i}\,\beta_1 + X_{2i}\,\beta_2 + \ldots + X_{ki}\,\beta_k \quad (7)$$

where $P(Y_i=1)$ is the probability that injury is present given a crash i and $x_1, x_2, \ldots x_k$ represents the crash characteristic variables and selected interaction terms. All categorical crash characteristics were transformed into one-hot encoded dummy variables before model estimation. Selected interaction terms were specified based on representative factor combinations identified from the retained association rules, allowing the logistic regression models to examine whether these co-occurring crash conditions were statistically associated with injury outcomes after controlling for individual crash characteristics. The models were estimated separately for each latent class using the glm() function in the base R stats package with a binomial distribution and logit link.

Result

Latent Class Analysis

Determining the optimal number of latent classes is a very important component of LCA. In this study, alternative LCA models were compared using several model selection criteria, including AIC, BIC, and CAIC, to identify the class structure that best represents the underlying heterogeneity in the crash data. The AIC, BIC, and CAIC values are plotted for models with 1 to 7 latent classes in Fig. 1. Based on Figure 1, the 2-class model exhibits the lowest BIC and CAIC value, indicating that it is the best fitting model among different number of latent classes. AIC instead continues to decrease as the number of latent classes increased. Since BIC and CAIC are asymptotically consistent and are widely considered reliable indicates of model fit in LCA (Labbo et al., 2024), we chose the 2-class model for subsequent interpretation and further analysis in this study and consider it as the best representation of the underlying heterogeneity. Therefore, in this study, we analyzed 2 latent classes representing two distinct latent classes of level 2 ADAS crashes, using ARM and logistic regression.

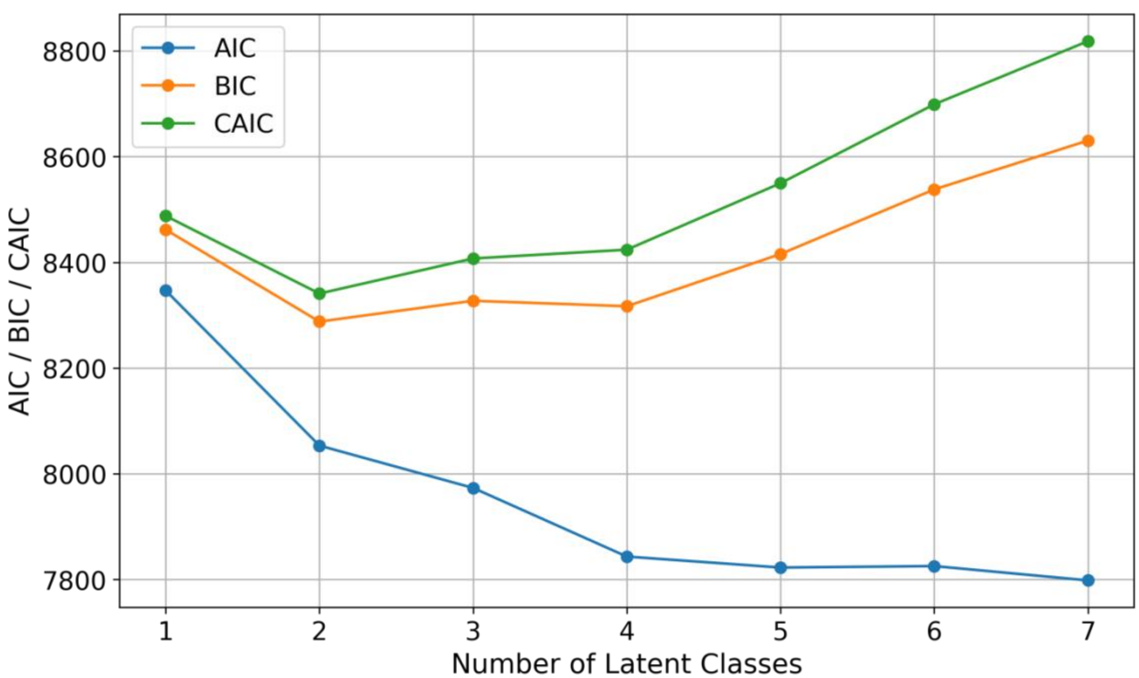

Fig. 1. AIC, BIC, and CAIC Values for Different Numbers of Latent Classes.

Table 2 presents the percentage distribution of crash characteristics across the two latent classes. Latent class 0 contained 217 crashes, while latent class 1 contained 400 crashes.

1) Latent Class 0: Lower-Speed Local Road Crashes

Latent class 0 represents a lower-speed, local road crash pattern. In this latent class, 50.6% of crashes had a pre-crash speed below 20 mph, and no crashes had a pre-crash speed above 60 mph. Crashes in this class were more commonly associated with streets and intersections, which together accounted for 72.4% of cases. Compared with latent class 1, latent class 0 shows more diverse subject vehicle pre-crash movements and contact area, suggesting that this class captures a broader set of lower-speed maneuvering and interaction scenarios more common in local road environments.

2) Latent Class 1: High-Speed Highway Crashes

Latent class 1 represents a higher-speed crash pattern. In this latent class, 66.2% of crashes have a pre-crash speed above 60 mph, and 89.9% occurred on highways or freeways. The subject vehicle was overwhelmingly moving straight or passing before the crash, accounting for 94.8% of cases. This latent class was also dominated by frontal impacts of the subject vehicle, with 73.3% of crashes involving front contact. Overall, latent class 1 appears to capture high-speed through-movement crashes in highway environments, where crashes are more likely to involve forward travel, front-end impacts, and less diverse pre-crash maneuvers.

Table 2. Characteristics of ADAS-Involved Crashes in Every Latent Class

Features		Latent Class 0 (%)	Latent Class 1 (%)	Overall (%)
Speed (MPH)	0-20	50.6	0	17.7
	20-40	32.7	7.3	16.2
	40-60	16.7	26.5	23.1
	>60	0	66.2	43.1
Weather Condition	Clear	78.7	70.3	73
	Cloudy	19.7	16	17.1
	Adverse	1.6	13.7	9.9

Time of Day	Morning	15.3	30.9	25.6
	Afternoon	32.4	27.6	29.2
	Evening	28.8	23	25
	Late Night	23.5	18.5	20.2
Roadway Type	Highway / Freeway	21.2	89.9	66.9
	Street	40	5	16.7
	Intersection	32.4	0.3	11
	Parking Lot	6.5	0.3	2.4
	Rural Road	0	4.4	3
Roadway Surface	Dry	89.3	85.9	87
	Wet / Snow / Slush / Ice	10.7	14.1	13
Subject Vehicle Pre-Crash Movement	Moving Straight or Passing	56.4	94.8	83.2
	Lane Changes and Merging	0.8	4.9	3.6
	Turning	23.4	0.4	7.3
	Stopped/Parked/ Backing	19.4	0	5.8
Crash Partner	Light-to-Medium Duty Vehicle	47.2	25.7	32.6
	Heavy Duty Vehicle	4.2	10.5	8.4
	Non-Motorist & Motorcycle	12.7	2.7	5.9
	Fixed Object	35.9	51	46.1
	Animal	0	10.1	6.8
	Front	36.9	73.3	60.9

Subject Vehicle Contact Area	Side	19.8	16.8	17.8
	Rear	28	7.3	14.4
	Other	15.3	2.6	7
Subject Vehicle Mileage Driven	0-25,000	65.4	54.6	58.3
	25,000-50,000	21.3	26.2	24.5
	>50,000	13.4	19.2	17.2
Subject Vehicle Model Year	Before 2020	17	12	13.8
	In/After 2020	83	88	86.2
Injury Classification	No Injuries Reported	53	49.2	50.6
	Minor/Moderate Injuries Reported	30.9	29.5	30
	Serious Injuries/Fatality Reported	16.1	21.2	19.4

Association Rule Mining

As introduced in Section 3.3, a two-stage robustness analysis combining threshold sensitivity analysis and bootstrap resampling was conducted within each latent class to identify association rules that were both interpretable and reproducible. Specifically, the minimum support thresholds were varied across 0.03, 0.04, and 0.05, while the minimum confidence thresholds were varied across 0.50, 0.55, and 0.60, resulting in nine support-confidence threshold combinations. To control rule complexity and avoid combinatorial explosion, the maximum length of itemsets was restricted to five items. For ARM analysis, injury severity was binarized into injury versus no injury, and only rules with injury as the consequent were retained. Under each support-confidence threshold combination, a rule was considered reproducible only if it reappeared in at least 80% of the bootstrap samples. For the final analysis, we retained only rules that satisfied this bootstrap reproducibility criterion in at least six of the nine threshold combinations. For each retained rule, we report its antecedent, the support count, support, confidence, lift calculated from the original dataset, and the number of threshold combinations in which the rule satisfied the bootstrap reproducibility criterion (referred to as selection frequency). A rule coverage table detailing the proportion of injury crashes explained by the retained rules can be found in Appendix A.

Association Rules for Latent Class 0

Latent class 0 represents a lower-speed, local road crash pattern. Table 3 presents the retained association rules ranked by lift, with injury outcome as the consequent in latent class 0. Overall, the results suggest that injury outcomes in this latent class were mainly associated with frontal impact configurations, intersection-related crashes, moderate-to-high speed ranges, and their combinations with roadway surface, crash partner, vehicle model year, and time-of-day characteristics. The highest-lift rule was subject vehicle contact area = front and speed = 40–60 mph with a confidence of 1.000 and lift of 2.127, suggesting that this combination was strongly associated with injury outcomes within latent class 0. The combination of roadway type = intersection and subject vehicle contact area = front appeared in all nine threshold combinations, with a confidence of 0.944 and a lift of 2.009, suggesting a highly stable and strong injury-related pattern. Similarly, roadway surface = dry and roadway type = intersection was retained across all threshold combinations, indicating that injury-related intersection crashes in this class were not limited to adverse roadway surface conditions. The single-factor rules including speed = 40–60 mph, subject vehicle contact area = front, and roadway type = intersection were also retained, but their lift values are lower than those of the combined rules. This suggests that although these individual factors were associated with injury outcomes, their combinations capture more specific and stronger injury-related crash scenarios. Overall, the results indicate that injury occurrence in latent class 0 were most strongly linked to frontal impacts at intersections and moderate-speed crash configurations.

Table 3. Retained Association Rules With Injury Outcome as the Consequent in Latent Class 0

Antecedent	Support	Confidence	Lift	Selection Frequency	Support Count
Subject Vehicle Contact Area=Front, Speed (MPH)=40-60	0.06	1	2.127	6	13
Roadway Type=Intersection, Subject Vehicle Contact Area=Front	0.078	0.944	2.009	9	17
Roadway Surface=Dry, Speed (MPH)=40-60	0.06	0.867	1.844	6	13
Subject Vehicle Contact Area=Front, Speed (MPH)=20-40	0.06	0.812	1.729	6	13

Roadway Surface=Dry, Roadway Type=Intersection	0.088	0.792	1.684	9	19
Crash Partner=Light-to-Medium Duty Vehicle, Roadway Type=Intersection	0.06	0.765	1.627	6	13
Speed (MPH)=40-60	0.092	0.741	1.576	9	20
Subject Vehicle Contact Area=Front, Subject Vehicle Model Year=In/After 2020	0.171	0.712	1.514	9	37
Subject Vehicle Contact Area=Front	0.175	0.655	1.394	8	38
Time of Day=Morning	0.078	0.654	1.391	6	17
Roadway Type=Intersection	0.161	0.636	1.354	6	35

Association Rules for Latent Class 1

Latent class 1 represents a higher-speed crash pattern. Table 4 presents the retained association rules ranked by lift, with injury outcome as the consequent in latent class 1. Overall, the retained rules suggest that injury outcomes in this latent class were mainly associated with high-speed crashes, frontal impacts, fixed-object collisions, light-to-medium duty vehicle crash partners, and specific temporal or environmental contexts such as morning, late night, and clear weather conditions. The highest-lift rule was crash partner = fixed object, subject vehicle contact area = front, subject vehicle pre-crash movement = moving straight or passing, and weather condition = clear, with a confidence of 0.774 and a lift of 1.526. However, because this rule had a support of 0.060 and appeared only in six of the nine threshold combinations, it should be interpreted as a relatively specific but meaningful injury-related pattern. High-speed conditions were especially prominent in this latent class. The rule crash partner = light-to-medium duty vehicle and speed > 60 mph appeared in all nine threshold combinations, with a confidence of 0.738 and a lift of 1.454, indicating a robust association between high-speed vehicle-to-vehicle crashes and injury outcomes. Similarly, crash partner = fixed object, subject vehicle contact area = front, and speed > 60 mph was retained across all nine threshold combinations, with a confidence of 0.717 and a lift of 1.414. These results suggest that high-speed impacts, particularly when combined with frontal contact or fixed-object collisions, are central injury-related patterns in latent class 1. Several additional rules provide contextual support. The rule speed > 60 mph with morning and clear weather indicates that high-speed injury-related crashes may occur under normal visibility or non-adverse weather conditions, rather than only under poor environmental conditions. The rule subject vehicle contact area = front and time of day = late night also suggests that late-night front-contact crashes are associated with elevated injury occurrence.

Table 4. Retained Association Rules With Injury Outcome as the Consequent in Latent Class 1

Antecedent	Support	Confidence	Lift	Selection Frequency	Support Count
Crash Partner=Fixed Object, Subject Vehicle Contact Area=Front, Subject Vehicle Pre-Crash Movement = Moving Straight or Passing, Weather Condition=Clear	0.06	0.774	1.526	6	24
Crash Partner=Light-to-Medium Duty Vehicle, Speed (MPH)=>60	0.078	0.738	1.454	9	31
Speed (MPH)=>60, Time of Day=Morning, Weather Condition=Clear	0.06	0.727	1.433	6	24
Crash Partner=Fixed Object, Subject Vehicle Contact Area=Front, Speed (MPH)=>60	0.082	0.717	1.414	9	33
Subject Vehicle Contact Area=Front, Time of Day=Late Night	0.06	0.706	1.391	7	24
Subject Vehicle Mileage Driven=>50,000	0.078	0.674	1.328	9	31
Speed (MPH)=>60, Time of Day=Morning	0.095	0.633	1.248	6	38
Crash Partner=Light-to-Medium Duty Vehicle	0.12	0.632	1.244	6	48

Subject Vehicle Contact Area=Front, Speed (MPH)=>60	0.205	0.631	1.243	6	82
Subject Vehicle Model Year=Before 2020	0.075	0.625	1.232	6	30

According to Table 3 and Table 4, no identical association rule was retained in both latent classes. However, several antecedent items appeared across both classes, including front contact, light-to-medium duty vehicle crash partner, and morning crashes.

Logistic Regression

Within each latent class, a binary logistic regression model was developed to investigate contributing factors to injury outcome of level 2 ADAS crashes. The no-injury crash was used as the reference level of dependent variable. The one-hot encoded crash characteristics listed in Table 1 were used as independent variables. In addition, selected interaction terms corresponding to the retained association rules were included. The interaction terms were selected based on relatively high lift values and high selection frequency across threshold combinations. QR decomposition was applied to identify and remove linearly dependent variables before estimation. The results were reported using parameter estimates, Wald z-statistics, p-values, and significance levels, where *, **, and *** indicate statistical significance at the 90%, 95%, and 99% confidence levels, respectively. Model-level statistics, including the number of estimated parameters, log-likelihood, were also reported.

Logistic Regression for Latent Class 0

The estimation results of the binary logistic regression model for latent class 0 are presented in Table 5. The model includes 37 estimated parameters and has a log-likelihood of -98.285. Several variables showed statistically significant associations with injury outcomes. The pre-crash speed category of 40–60 mph is positively associated with injury outcomes at the 90% confidence level. This finding is generally consistent with the ARM results, which also identified speed-related rules as injury-associated patterns. The results also show that subject vehicle contact area and vehicle mileage were important predictors. The coefficients for front contact area and other contact area are negative and statistically significant, indicating that these categories were associated with lower injury odds when considered as main effects. Similarly, all three subject vehicle mileage categories are negatively associated with injury outcomes, with the strongest statistical significance observed for the 0–25,000 and 25,000–50,000 mileage groups. In contrast, vehicles with model years before 2020 were positively associated with injury outcomes at the 95% confidence level. Among the rule-based interaction terms, the interaction between front contact area and pre-crash speed of 40–60 mph is not statistically significant in the model. However, the coefficient of interaction between front contact area and model year in/after 2020 is positive and statistically significant at the 95% confidence level.

Table 5. Estimation Results of Binary Logistic Regression Model of Latent Class 0

Variables		Parameter Estimate	Z value	P value
(Intercept)		0.045	0.063	0.95
Speed (MPH)	0-20	-0.213	-0.323	0.747
	20-40	0.644	0.863	0.388
	40-60	1.605*	1.773	0.076
Weather Condition	Clear	0.099	0.13	0.897
	Cloudy	0.098	0.098	0.921
	Adverse	18.063	0.005	0.996
Subject Vehicle Pre-Crash Movement	Moving Straight or Passing	-0.919	-1.607	0.108
	Lane Changes and Merging	18.73	0.003	0.998
	Stopped/Parked/ Backing	-0.69	-0.937	0.349
	Turning	-0.592	-0.874	0.382
Crash Partner	Fixed Object	-0.707	-1.049	0.294
	Light-to-Medium Duty Vehicle	-0.116	-0.234	0.815
	Non-Motorist & Motorcycle	0.932	1.243	0.214
	Heavy Duty Vehicle	1.4	0.885	0.376
Roadway Type	Highway / Freeway	1.18	1.462	0.144

	Intersection	1.148	1.412	0.158
	Parking Lot	-0.332	-0.285	0.775
	Street	0.242	0.329	0.742
Roadway Surface	Dry	0.454	0.508	0.611
	Wet / Snow / Slush / Ice	1.513	1.133	0.257
Time of Day	Afternoon	-0.481	-0.621	0.535
	Evening	-0.007	-0.009	0.993
	Late Night	0.06	0.082	0.934
	Morning	0.737	0.972	0.331
Subject Vehicle Contact Area	Front	-2.647*	-1.698	0.089
	Other	-2.659**	-2.055	0.04
	Rear	-0.335	-0.547	0.584
	Side	-0.871	-1.396	0.163
Subject Vehicle Mileage Driven	>50,000	-1.976**	-2.253	0.024
	0-25,000	-1.514***	-3.015	0.003
	25,000-50,000	-2.207***	-3.1	0.002
Subject Vehicle Model Year	Before 2020	1.258**	2.125	0.034
(Subject Vehicle Contact Area = Front) + (Speed (MPH) = 40-60)		16.895	0.012	0.99
(Roadway Type = Intersection) + (Subject Vehicle Contact Area = Front)		1.585	1.09	0.276
(Roadway Surface = Dry) + (Roadway Type= Intersection)		0.442	0.412	0.681

(Subject Vehicle Contact Area = Front) + (Subject Vehicle Model Year=After 2020)	3.688**	2.364	0.018
Number of Parameters	37		
Loglikelihood	-98.285		

Note: *P < 0.1; ** P < 0.05; ***P < 0.01.

Logistic Regression for Latent Class 1

The estimation results of the binary logistic regression model for latent class 1 are presented in Table 6. The model includes 38 estimated parameters and has a log-likelihood of -231.528. Several main-effect variables and rule-based interaction terms are statistically significant. Among the main-effect variables, cloudy weather is negatively associated with injury outcomes at the 99% confidence level. Lane changes and merging before the crash is positively associated with injury outcomes at the 90% confidence level. For crash partner, animal-related crashes were negatively associated with injury outcomes at the 95% confidence level, while the heavy-duty vehicle crash-partner category is positively associated with injury outcomes at the 90% confidence level. Rural road crash location is positively associated with injury outcomes at the 90% confidence level. Evening crash time is positively associated with injury outcomes at the 95% confidence level. Subject vehicle mileage driven between 0 and 25,000 miles is negatively associated with injury outcomes at the 90% confidence level. Among the rule-based interaction terms, the interaction between light-to-medium duty vehicle crash partners and speeds above 60 mph is positively associated with injury outcomes at the 90% confidence level. The interaction among fixed-object crash partners, front contact area, and speeds above 60 mph is positively associated with injury outcomes at the 95% confidence level. The interaction between front contact area and late-night crashes is positively associated with injury outcomes at the 99% confidence level. These three interaction terms were consistent with the ARM results for latent class 1.

Table 6. Estimation Results of Binary Logistic Regression Model of Latent Class 1

Variables		Parameter Estimate	Z value	P value
(Intercept)		13.183	0.015	0.988
Speed (MPH)	>60	-0.134	-0.347	0.728
	20-40	-0.945	-1.521	0.128

	40-60	-0.053	-0.14	0.889
Weather Condition	Adverse	-0.489	-0.602	0.547
	Clear	-0.504	-1.284	0.199
	Cloudy	-1.404***	-2.638	0.008
Subject Vehicle Pre-Crash Movement	Lane Changes and Merging	1.394*	1.859	0.063
	Moving Straight or Passing	0.136	0.414	0.679
	Turning	-13.828	-0.016	0.988
Crash Partner	Animal	-1.199**	-2.177	0.029
	Fixed Object	-0.204	-0.599	0.549
	Heavy Duty Vehicle	0.905*	1.85	0.064
	Light-to-Medium Duty Vehicle	0.136	0.287	0.774
	Non-Motorist & Motorcycle	1.696	1.503	0.133
Roadway Type	Highway / Freeway	-0.323	-0.796	0.426
	Other	-0.967	-0.612	0.54
	Rural Road	1.326*	1.732	0.083
	Street	-0.773	-1.098	0.272
Roadway Surface	Dry	0.2	0.532	0.595
	Wet / Snow / Slush / Ice	0.295	0.364	0.716
Time of Day	Afternoon	0.348	0.793	0.428
	Evening	0.913**	2.106	0.035
	Late Night	-0.353	-0.605	0.545

	Morning	0.534	1.038	0.299
Subject Vehicle Contact Area	Front	-0.044	-0.12	0.905
	Other	-0.487	-0.574	0.566
	Rear	-0.525	-0.908	0.364
	Side	-0.327	-0.755	0.45
Subject Vehicle Mileage Driven	>50,000	0.649	1.507	0.132
	0-25,000	-0.49*	-1.739	0.082
	25,000-50,000	-0.187	-0.531	0.595
Subject Vehicle Model Year	Before 2020	-12.769	-0.014	0.989
	In/After 2020	-13.26	-0.015	0.988
(Crash Partner= Light-to-Medium Duty Vehicle) + (Speed (MPH) = >60)		1.01*	1.654	0.098
(Crash Partner= Fixed Object) + (Subject Vehicle Contact Area = front) + (Speed (MPH) = >60)		1.14**	2.174	0.03
(Subject Vehicle Contact Area = front) + (Time of Day= Late Night)		2.192***	3.089	0.002
Number of Parameters		38		
Loglikelihood		-231.528		

Note: *P < 0.1; ** P < 0.05; ***P < 0.01.

Discussion

To gain a deeper understanding of crash scenarios that are associated with higher injury likelihood in crashes involving Level 2 ADAS-equipped vehicles, this section provides an in-depth discussion of key risk factors and representative crash patterns identified across distinct latent classes derived from latent class analysis. Specifically, the identified latent classes capture heterogeneous injury-prone scenarios, including: latent class 0: Lower-speed local road crashes and latent class 1: High-speed highway crashes.

Building on the findings of Esmaili et al. (2023) who reported that subject vehicle front impact was associated with injury occurrence and Ding et al. (2024) who reported that frontal impact to the subject vehicle was associated with a higher predicted probability of moderate-to-severe injury in ADAS crashes, our analysis suggests that the role of front contact area should be interpreted in relation to the broader crash context rather than as an isolated factor. In the ARM results, front contact area frequently appeared in injury-related rules with lift values greater than 1, when combined with other crash characteristics. However, the logistic regression results indicate that frontal contact area alone is not consistently associated with a positive effect on injury occurrence. In latent class 0, front contact area has a statistically significant negative coefficient, while its interaction with subject vehicle model year in/after 2020 has a positive and statistically significant coefficient. In latent class 1, front contact area alone is not statistically significant, but the interactions involving front contact area with fixed-object crashes at speeds above 60 mph and with late-night crashes are both positively and significantly associated with injury outcomes. Overall, these results suggest that frontal impact is better understood as a context-dependent crash characteristic associated with injury occurrence. Its association with injury outcomes becomes more evident when frontal contact occurs in combination with specific crash contexts, such as particular vehicle model-year groups, high-speed fixed-object crashes, or late-night conditions.

Real-world and track-based evidence suggests important performance gaps under low-light conditions, with pedestrian automatic emergency braking showing little benefit in darkness and unlit roadway environments (IIHS,2022). This limitation may help explain why frontal impacts combined with late-night conditions were associated with injury outcomes in latent class 1. However, this explanation remains speculative, because the present study does not directly measure ADAS activation, system performance, or causal crash-avoidance mechanisms.

Extending the findings of Esmaili et al. (2023) who identified a correlation between injury outcome and high pre-crash speed and Tamakloe et al. (2025) who reported that crash groups with a higher frequency of injury outcomes often involved higher pre-crash speeds, our results further show that pre-crash speed is an important injury-related factor, but its role varies by latent class. In the ARM results, speed-related factors frequently appear in retained rules with lift values greater than 1, particularly when combined with other crash characteristics. In latent class 0, crashes with pre-crash speeds of 40–60 mph are positively associated with injury outcomes in the logistic regression model, and speed-related rules involving 40–60 mph also appear in the ARM results.

In contrast, in latent class 1, speeds above 60 mph are mainly associated with injury outcomes through interaction terms rather than as a standalone main effect. Specifically, high-speed crashes become statistically significant when combined with light-to-medium duty vehicle crash partners or fixed-object crashes with front contact area. From a practical perspective, although this study does not establish causal effects, the observed associations highlight the importance of incorporating higher-speeds and context-dependent crash configurations into ADAS evaluation and validation. In particular, scenario-based testing could place greater emphasis on combinations involving moderate and high pre-crash speeds, frontal contact, fixed-object crashes, light-to-medium duty vehicle crash partners, and late-night conditions. These findings may help manufacturers refine speed-aware system-use guidance, driver warnings, and operational design constraints for Level 2 ADAS-equipped vehicles.

Several other variables also warrant discussion. Prior studies have reported that crashes involving heavy-duty vehicles are associated with a higher likelihood of injury outcomes (Fu et al., 2025; Esmaili et al., 2023). In our analysis, this pattern is observed only in latent class 1, where heavy-duty vehicle crash partners were positively and statistically significantly associated with injury outcomes. This suggests that the injury relevance of heavy-duty vehicle involvement may depend on the broader crash context represented by the latent class structure. Intersection-related crashes have also been reported as injury-relevant in previous studies (Esmaili et al., 2023; Ding et al., 2024). In addition, Tamakloe et al. (2025) identified the largest ADAS-related crash cluster as being dominated by fatal left-turn crashes at intersections. In our results, intersection-related patterns are prominent in latent class 0, where several association rules involving intersections have lift values greater than 1. However, the intersection variable and the intersection-related interaction terms are not statistically significant in the logistic regression model. This discrepancy may arise because ARM identifies overrepresented attribute combinations, while logistic regression evaluates whether these variables provide additional explanatory power after controlling for other crash characteristics.

The analysis also shows that vehicle model year has opposite associations with injury across latent classes. In latent class 0, ARM identifies an injury-related rule that includes model year in/after 2020 in combination with frontal contact, indicating a conditional association within a specific crash context. In latent class 1, model year before 2020 appears as a standalone antecedent, which suggests a more general association with injury crashes within the latent class. These patterns may reflect differences in crashworthiness and vehicle safety design standards over time (e.g., model year vehicles before 2020 were developed under different side crash testing protocols (IIHS, 2026)) or differences in model year exposure across different crash environments. However, because the hybrid LCA+ARM framework does not allow for causal interpretation, no definitive conclusions can be drawn about the reasons for these differences.

Overall, our framework provides transparent and interpretable insights into injury risk patterns in crashes involving Level 2 ADAS-equipped vehicles. By first applying LCA to account for unobserved heterogeneity in crash scenarios and then using ARM to uncover high-risk variable combinations within each latent class and using logistic regression, this workflow enables the identification of context-specific association that would be obscured in aggregated analyses. The proposed hybrid and interpretable approach is readily transferable to other traffic safety applications, where understanding heterogeneous risk patterns and interacting factors is critical for informing targeted policy interventions and system design improvements.

Exploratory Electric Vehicle Crash Assessment

Using Pennsylvania Department of Transportation (PennDOT) crash data from 2022-2024, we conduct an exploratory analysis of contributing factors in electric vehicle (EV) crashes in Pennsylvania. Based on Table 7, the majority of EV crashes occurred at intersections or on local streets, in suburban settings, at posted speed limits between 20 and 40 mph, during daylight, and under dry roadway conditions. The clearest severity gradient appears in collision type: head-on collisions and pedestrian strikes together account for 27.4% of crashes with serious injuries or fatalities but less than 3% of no-injury crashes, and single-vehicle crashes are over-represented among serious injury and fatality crashes (29.6%) relative to minor or moderate injury crashes (13.9%). Based on Table 8, alcohol involvement shows the steepest gradient across severity levels, recorded in 3.9% of no-injury crashes, 4.7% of minor or moderate injury crashes, and 16.3% of crashes with serious injuries or fatalities. Running a red light (4.5% to 7.4%) and mature EV drivers (10.1% to 14.8%) also increase with severity, whereas aggressive driving shows no consistent gradient. In contrast, the BEV share is nearly constant across the three severity levels (67.3%, 67.1%, and 71.1%), indicating that powertrain type is not differentiated by crash severity.

Table 7. Crash Context Characteristics of EV Crashes Used for the Analysis, by Injury Severity

Feature	Category	No Injuries Reported	Minor/Moderate Injuries Reported	Serious Injuries/Fatality Reported
Collision Type	Angle	774 (35.2%)	749 (44.3%)	48 (35.6%)
	Rear-End	594 (27.0%)	483 (28.6%)	20 (14.8%)
	Hit Fixed Object	429 (19.5%)	170 (10.1%)	20 (14.8%)
	Sideswipe (Same Direction)	166 (7.6%)	95 (5.6%)	2 (1.5%)
	Head-On	61 (2.8%)	61 (3.6%)	19 (14.1%)
	Other/Unknown	107 (4.9%)	13 (0.8%)	1 (0.7%)
	Hit Pedestrian	1 (0.0%)	68 (4.0%)	18 (13.3%)
	Sideswipe (Opposite Direction)	41 (1.9%)	36 (2.1%)	3 (2.2%)
	Non-Collision	19 (0.9%)	11 (0.7%)	4 (3.0%)

Feature	Category	No Injuries Reported	Minor/Moderate Injuries Reported	Serious Injuries/Fatality Reported
	Rear-to-Rear	6 (0.3%)	4 (0.2%)	0 (0.0%)
Vehicle Involvement	Multi Vehicle	1684 (76.6%)	1455 (86.1%)	95 (70.4%)
	Single Vehicle	514 (23.4%)	235 (13.9%)	40 (29.6%)
Roadway Type	Intersection	815 (37.1%)	770 (45.6%)	61 (45.2%)
	Street/Other	709 (32.3%)	517 (30.6%)	51 (37.8%)
	Highway/Freeway	563 (25.6%)	336 (19.9%)	20 (14.8%)
	Ramp	111 (5.1%)	67 (4.0%)	3 (2.2%)
Urban/Rural	Suburban	1761 (80.1%)	1430 (84.6%)	105 (77.8%)
	Rural	437 (19.9%)	260 (15.4%)	30 (22.2%)
Speed (MPH)	20-40	1160 (52.8%)	956 (56.6%)	80 (59.3%)
	40-60	787 (35.8%)	606 (35.9%)	43 (31.9%)
	>60	160 (7.3%)	46 (2.7%)	7 (5.2%)
	0-20	91 (4.1%)	82 (4.9%)	5 (3.7%)
Lighting	Daylight/Dawn/Dusk	1536 (69.9%)	1238 (73.3%)	93 (68.9%)
	Dark	662 (30.1%)	452 (26.7%)	42 (31.1%)
Weather	Dry/Clear	1704 (77.5%)	1396 (82.6%)	111 (82.2%)
	Wet	423 (19.2%)	265 (15.7%)	23 (17.0%)
	Icy/Snow	71 (3.2%)	29 (1.7%)	1 (0.7%)
Roadway Condition	No Unusual Conditions	2152 (97.9%)	1671 (98.9%)	134 (99.3%)
	Work Zone	46 (2.1%)	19 (1.1%)	1 (0.7%)
Season	Fall	641 (29.2%)	493 (29.2%)	27 (20.0%)
	Summer	549 (25.0%)	460 (27.2%)	45 (33.3%)
	Winter	509 (23.2%)	373 (22.1%)	27 (20.0%)
	Spring	499 (22.7%)	364 (21.5%)	36 (26.7%)
Model Year	In/After 2020	1679 (76.4%)	1334 (78.9%)	104 (77.0%)
	Before 2020	519 (23.6%)	356 (21.1%)	31 (23.0%)

Note: model year refers to the newest vehicle involved in the crash.

Table 8. Driver Behavior and Powertrain Characteristics of EV Crashes Used for the Analysis, by Injury Severity

Feature	Category	No Injuries Reported	Minor/Moderate Injuries Reported	Serious Injuries/Fatality Reported
Driver Behavior	Aggressive Driving	1424 (64.8%)	1178 (69.7%)	84 (62.2%)
	Alcohol Involvement	85 (3.9%)	79 (4.7%)	22 (16.3%)
	Running Red Light	99 (4.5%)	114 (6.7%)	10 (7.4%)
EV Driver Age	Young Driver (age ≤20)	172 (7.8%)	81 (4.8%)	10 (7.4%)
	Mature Driver (age ≥65)	221 (10.1%)	179 (10.6%)	20 (14.8%)
Powertrain	BEV	1479 (67.3%)	1134 (67.1%)	96 (71.1%)
	PHEV	719 (32.7%)	556 (32.9%)	39 (28.9%)
	ICEV Involved	1590 (72.3%)	1430 (84.6%)	107 (79.3%)

Note: behavior flags are crash-level (any involved driver); driver age refers to the EV's own driver; "ICEV Involved" indicates crashes that also involve at least one internal combustion engine vehicle. Categories within a feature are not mutually exclusive for this table.

Project Outputs

- In November 2025, project PI Corey Harper participated in the Safety21 Deployment Partner Consortium Symposium.
- In January 2026, project PI Corey Harper presented a paper at the Transportation Research Board 105th Annual Meeting titled "Improving our understanding of contributing factors in California level 4 autonomous vehicle crashes through latent class analysis and machine learning-SHAP interpretations"
- PI Corey Harper and Co-PI Chris Hendrickson published a paper on Level 4 AV Crash Safety: Wang, J., Harper, C. D., & Hendrickson, C. (2026). Improving our understanding of contributing factors in California level 4 autonomous vehicle crashes through latent class analysis and machine learning-SHAP interpretations. *Journal of Safety Research*, *96*, 19-37.
- PI Corey Harper and Co-PI Chris Hendrickson published a paper on Level ADAS Crash Safety: Wang, J., Harper, C. D., & Hendrickson, C. (2026). A hybrid latent class analysis and association rule mining framework to identify injury-related risk patterns in level 2 advanced driver assistance system crashes. *Accident Analysis & Prevention*, *235*, 108631.

Reference

SAE, 2021. SAE Levels of Driving Automation Refined for Clarity and International Audience.https://www.sae.org/news/blog/sae-levels-driving-automation-clarity-refinements (accessed 1.1.26).

Khan, A., Harper, C.D., Hendrickson, C.T., Samaras, C., 2019. Net-societal and net-private benefits of some existing vehicle crash avoidance technologies. Accident Analysis & Prevention 125, 207–216. doi:10.1016/j.aap.2019.02.003

Harper, C.D., Hendrickson, C.T., Samaras, C., 2016. Cost and benefit estimates of partially-automated vehicle collision avoidance technologies. Accident Analysis & Prevention 95, 104–115. doi:10.1016/j.aap.2016.06.017

NHTSA, 2021. Standing General Order on Crash Reporting. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting (accessed 2.2.26).

Wang, J., Harper, C.D., Hendrickson, C., 2026. Improving our understanding of contributing factors in California level 4 autonomous vehicle crashes through latent class analysis and machine learning-SHAP interpretations. Journal of Safety Research 96, 19–37. doi:10.1016/j.jsr.2025.11.007

Morin, S., Legault, R., Laliberté, F., Bakk, Z., Giguère, C.-É., De La Sablonnière, R., Lacourse, É., 2025. StepMix : A Python Package for Pseudo-Likelihood Estimation of Generalized Mixture Models with External Variables. J. Stat. Soft. 113 8 . doi:10.18637/jss.v113.i08

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 6 , 716–723. doi:10.1109/TAC.1974.1100705

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Statist. 6 2 . doi:10.1214/aos/1176344136

Weller, B.E., Bowen, N.K., Faubert, S.J., 2020. Latent Class Analysis: A Guide to Best Practice. Journal of Black Psychology 46 4 , 287–311. doi:10.1177/0095798420930932

Padgett, R.N., Tipton, R.J., 2020. Identifying latent classes with ordered categorical indicators. doi:10.48550/ARXIV.2009.07345

Labbo, M.S., Qu, L., Xu, C., Bai, W., Ayele Atumo, E., Jiang, X., 2024. Understanding risky driving behaviors among young novice drivers in Nigeria: A latent class analysis coupled with association rule mining approach. Accident Analysis & Prevention 200, 107557. doi:10.1016/j.aap.2024.107557

Liu, Pei, Guo, Y., Liu, Pan, Ding, H., Cao, J., Zhou, J., Feng, Z., 2024. What can we learn from the AV crashes? – An association rule analysis for identifying the contributing risky factors. Accident Analysis & Prevention 199, 107492. doi:10.1016/j.aap.2024.107492

Fagerlind, H., Harvey, L., Humburg, P., Davidsson, J., Brown, J., 2022. Identifying individual-based injury patterns in multi-trauma road users by using an association rule mining method. Accident Analysis & Prevention 164, 106479. doi:10.1016/j.aap.2021.106479

Hossain, A., Sun, X., Thapa, R., Codjoe, J., 2022. Applying Association Rules Mining to Investigate Pedestrian Fatal and Injury Crash Patterns Under Different Lighting Conditions. Transportation Research Record: Journal of the Transportation Research Board 2676 6 , 659‑672. doi:10.1177/03611981221076120

Tamakloe, R., Khorasani, M., Das, S., Kim, I., 2025. Pattern recognition in crash clusters involving vehicles with advanced driving technologies. Accident Analysis & Prevention 218, 108072. doi:10.1016/j.aap.2025.108072

Jamison, L., 2024. Bootstrapped Association Rule Mining in R [WWW Document]. URL https://library.virginia.edu/data/articles/bootstrapped-association-rule-mining-r

Esmaili, A., Oshanreh, M., Regan, A., 2023. Investigation of Crash Severities involving ADAS Level 2 and ADS-equipped vehicles. Presented at the Transportation Research Board

Ding, S., Abdel-Aty, M., Barbour, N., Wang, D., Wang, Z., Zheng, O., 2024. Exploratory analysis of injury severity under different levels of driving automation (SAE Levels 2 and 4) using multi-source data. Accident Analysis & Prevention 206, 107692. doi:10.1016/j.aap.2024.107692

IIHS, 2022. Pedestrian crash avoidance systems cut crashes — but not in the dark. https://www.iihs.org/news/detail/pedestrian-crash-avoidance-systems-cut-crashes--but-not-in-the-dark (accessed 9.10.24).

Fu, H., Ye, S., Fu, X., Chen, T., Zhao, J., 2025. New insights into factors affecting the severity of autonomous vehicle crashes from two sources of AV incident records. Travel Behaviour and Society 38, 100934. doi:10.1016/j.tbs.2024.100934

IIHS, 2026. Side. URL https://www.iihs.org/ratings/about-our-tests/side