Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 4323816, 16 pages
https://doi.org/10.1155/2020/4323816

Research Article

Discovering Insightful Rules among Truck Crash Characteristics using Apriori Algorithm

Jungyeol Hong, Reuben Tamakloe, and Dongjoo Park

Department of Transportation Engineering, The University of Seoul, 163 Seoulsiripdae-ro Dongdaemun-gu, Seoul 02504, Republic of Korea

Correspondence should be addressed to Dongjoo Park; djpark@uos.ac.kr

Received 9 August 2019; Revised 22 October 2019; Accepted 10 December 2019; Published 16 January 2020

Guest Editor: Joyoung Lee

This study aims to discover hidden patterns and potential relationships in risk factors in freight truck crash data. Existing studies mainly used parametric models to analyze the causes of freight vehicle crashes. However, predetermined assumptions and underlying relationships between independent and dependent variables have been cited as its limitations. To overcome these limitations and provide a better understanding of factors that lead to truck crashes on the expressways, we applied the Association Rules Mining (ARM) technique, which is a nonparametric method. ARM quantifies the interrelationships between the antecedents and consequents of truck-involved crashes and provides researchers with the most influential set of factors that leads to crashes. We utilized a freight vehicle-involved crash data consisting of 19,038 crashes that occurred on the Korean expressways from 2008 to 2017 for this investigation. From the data, 90,951 association rules were generated through ARM employing the Apriori algorithm. The lift values estimated by the Apriori algorithm showed the strength of association between risk factors, and based on the estimated lift values, we identified key crash contributory factors that lead to truck-involved crashes at various segment types, under different weather conditions, considering the driver's age, crash type, driver's faults, vehicle size, and roadway geometry type. From the generated rules, we demonstrated that overspeeding with medium-weight trucks was highly associated with crashes during the rainy weather, whereas drowsy driving during the evening was correlated with crashes during fine weather. Segment-related crashes were mainly associated with driver's faults and roadway geometry. Our results present useful insights and suggestions that can be used by transport stakeholders, including policymakers and researchers, to create relevant policies that will help reduce freight truck crashes on the expressways.

1. Introduction

The demand for freight transportation has been increasing concurrently with the growing population experienced globally [1]. In order to meet the needs of consumers, the number of trucks, which forms the most dominant modes of freight transportation, keeps on increasing [2]. Even though freight vehicles play crucial roles in the economies of countries, substantial volumes of truck traffic impose significant safety issues. While there may be fewer crashes involving freight trucks on expressways, the unique features of trucks with respect to their size and weight, and their operational characteristics contribute to the significant increase in fatalities and loss of property [1, 3]. As such, freight truck safety analysis continues to be a crucial topic in the transportation field [2].

Research on freight truck safety has focused on three main areas, namely, truck-involved crash frequency, severity in terms of damage caused and the degree of injury sustained by people involved in freight truck crashes, and the likelihood of truck-involved crash occurrence. In particular, driver characteristics, environmental factors, vehicle features, and geometric roadway design features have been cited as significant contributory of truck-involved crashes [2, 4]. To achieve valuable insights into freight truck safety, researchers employed a wide variety of methodologies, such as parametric discrete-outcome modeling techniques and nonparametric machine learning-based approaches. However, the majority of these studies typically focused on crash-risk factors and their contribution to each truck crash. Since every single crash is a result of a combination of interrelated

factors, they are to be studied thoroughly in order to find meaningful countermeasures for truck crashes [5].

Based on this recognition, recent studies have employed association rules mining (ARM), a data mining technique to identify interesting associations between contributory crash factors that simultaneously impact crashes [6]. To achieve this, the Apriori algorithm, which efficiently searches association rules using straightforward and easy to understand computations, has mainly been employed [7, 8]. Pande et al. [8] argue that this technique is more favorable compared to cluster analysis as it provides easily understood relationships between crash-risk factors. Also, it is better when compared to parametric methods because it does not require assigning variables as dependent variables or independent variables, and no predefined assumptions are underlying the relationships between dependent and independent variables. Hence, there are no assumptions to be violated to cause flawed estimation results [9].

Understanding truck crashes on the expressways requires full knowledge of the factors that combine to form those crashes. Expressways serve as essential links between cities, and freight truckers are among the majority that plies them. Due to the function of the expressways, they are characterized by unique features that contribute to increased crash severity and risk [10, 11]. The primary objective of this study is to identify patterns and relationships in expressway freight truck crash-risk factors that are usually unknown to researchers using ARM.

2. Literature Review

2.1. Factors Influencing Truck Crashes. Research has identified a wide variety of factors that influence truck-involved crash frequency. Cantor et al. [12] investigated the contribution of truck driver factors on the likelihood of truck-involved crashes using a driver-focused crash prediction model. Empirical results suggested that driver-specific contributory factors such as weight, gender, and height are related to the likelihood of crash occurrence. In line with these findings, Zhu et al. [13] also identified that male drivers had a higher chance of crash involvement as they often engage in risky driving on highways. Regarding vehicle maintenance, Cantor et al. [12] established that poorly maintained trucks increased crash probability. Detailed analysis by Dong et al. [14] demonstrated that the annual average daily traffic significantly affected truck-involved crash frequency and severity. Also, impaired drivers and driving under the influence of alcohol or drugs were observed to lead to a higher crash frequency. Furthermore, the authors identified that inclement weather conditions and intersection locations are linked with an increase in truck-involved crash frequency.

Studies considering driver's age has been characterized by inconsistencies in the literature [3]. Young drivers are associated with increased fatal and property damage only (PDO) crashes [15, 16]. Other studies also portrayed that younger drivers are more likely to have a decrease in no injury truck-involved crashes [17]. Chen and Chen [18] examined injury severities of truck drivers concentrating on single-vehicle and multivehicle crashes on highways. Their study reports that drivers older than 50 years increase the likelihood of fatal crashes in single-vehicle (SV) truck-involved crashes, whereas the probability decreases in multivehicle (MV) truck-involved crashes.

Several study outcomes have been consistent in the perspective of gender. Model estimations show that female truck drivers have an increased chance of severe and fatal injuries [17, 19, 20]. Using a highway truck crash data, Khattak et al. [21] also found that dangerous driving behaviors such as drunk driving significantly increases the injury severity risk of truck occupants. According to them, the severity of injuries also increases in trucks carrying hazardous materials. Speeding has been identified as one of the significant influencers of injury severity [2, 21, 22]. In the literature concerning occupant injury severity in rear-end truck-involved crashes, Yuan et al. [23] demonstrated that injury severity increases as speed increases. Also, Hao et al. [24] studied truck-involved crashes at highway-rail grade crossings in the US and found that injury severity decreases with the presence of speed control features on the roads.

Other factors such as inclement weather [23–25], wet road surface [15, 23], fatigued driving [24], poor visibility and dark conditions [17, 23, 24], the weight and number of vehicles involved in the crash [15], and truck driver at-fault crashes [3] increase injury severity probabilities. Concerning roadway features, Ahmed et al. [22] found that truck crash severity increased when trucks crash into fixed objects on the highway but reduce when they crash into guardrails. Results also confirmed that truck crash severity increases when it occurs on leveled road surfaces.

Park and Jovanis [26] investigated the influence of hours of service of truck drivers on truck crash risk. Results showed that crash risk increased from 50% to 260% compared to the first hour of driving. This result is in line with those observed by Teoh et al. [27]. They analyzed a matched case-control study using data comprising large truck-involved crashes from 2010 to 2012 and demonstrated that long hours of driving led to increased crash risk. Furthermore, their results showed that crash risk increased by up to three times when the truck had defects. Likewise, the truck crash probability showed to have a positive correlation with drivers' age and working conditions. Regarding roadway geometry factors, Yuan et al. [23] also noted that straight road sections of expressways were known for having increased probabilities of crashes.

2.2. Methodological Approaches Used in Freight Truck Crash Analyses. To date, a variety of techniques have been applied to understand the factors contributing to truck crashes. Researchers have broadly classified them into two, namely, parametric methods and nonparametric methods. Parametric models form the majority of models used in truck crash analysis. Among these, several variants of ordered logit and probit regression models are the most common. These models essentially help to determine the contribution of individual crash-risk factors to injury severities [15, 17, 18, 22, 23]. Although parametric models have served

well in research, analysts have criticized them for requiring a set of predetermined assumptions and underlying relationships between independent and dependent variables. Their prediction accuracy may be affected and turn out as low when these assumptions are violated [6, 28].

To overcome these challenges, some researchers have resorted to using nonparametric techniques for investigating traffic crash data. In particular, Lopez et al. [28] used decision rules to study the patterns of SV crashes on rural highways, de Oña et al. [29] employed decision trees to extract valuable information from police crash reports, and Kashani et al. [30] employed classification tree models to examine factors that affect injury severity of occupants of vehicles.

In other studies, researchers proposed a combination of parametric and nonparametric methods to achieve better predictions than either of the two methods would have given if used alone or separately [31]. In the field of road safety, Rusli et al. [32] used a similar technique to model crash severity. The authors argue that combining both approaches not only allows for the specification of nonlinearities and interactions but also the main effects. The parametric model used in their study allowed for capturing unobserved heterogeneity, which has been mentioned by previous research as being very important in accident analysis [10].

Nevertheless, some recent research studies have argued that since some nonparametric methods require a vast dataset for analysis, they may suffer from overfitting [6, 31]. Therefore, applying a nonparametric technique that can deal with a small number of variables while determining important patterns in crash data as in the association rules mining (ARM) approach is imperative. Also, unlike the parametric approaches, it requires no predefined underlying assumptions [6]. Its primary objective is to determine real associations in crash data without specifying dependent and independent variables [33]. Another advantage of an ARM is that it enables researchers to find easy and readily understandable causal relationships among interrelated factors in a crash database by way of good visualization [5].

Even though ARM is a popular nonparametric technique, only a few studies in the area of traffic safety have used it in their analyses. Das et al. [9] used it to identify patterns in traffic crashes under rainy weather conditions. Weng et al. [6] also used it to find patterns in work zone crashes. Yu et al. [31] analyzed 63,325 crashes from Wisconsin using ARM and identified that drivers are more inclined to having more severe crashes when the road surface is dry, and the weather is fine. They explained that drivers are more likely to indulge in risky driving under such good road surface and weather conditions.

Extending the ideas obtained from previous research, we primarily seek to analyze risk factors of truck-involved crashes that occurred on expressways in South Korea systematically using ARM and to identify significant associations between the truck-involved crash-risk factors and characteristics.

3. Data Description

Traffic crash records are collected and stored in the Korean Expressway Corporation (KEC) crash database system and managed by specially trained crash investigators from the KEC [34]. Thus, problems such as missing variables were absent. At the KEC, a reportable crash is one that causes either damage to property, injuries, or death of any person. Upon visiting the crash scene, the investigators give all vehicles involved in one crash a unique identification (ID) number irrespective of the fact that a driver was at fault or not. The association of a distinct ID number to each crash observation ensures that there are no duplicates in the data. The officials then proceed to collect crash-related information for each crash observation. The collected crash-related data consist of information pertaining to variables such as severity level, weather, vehicle characteristics, driver's age, time of day, roadway geometry, location, type of crash, and cause of the crash. The KEC classifies the severity of crashes into four, from A through D. Level A represents fatal crashes (all crashes where the number of deaths > 3, injured persons > 20, or property damage cost > 1 billion Korean Won KRW). Level B shows severe injury (represents all crashes where $1 <$ number of deaths ≤ 3, $5 <$ injured persons ≤ 20 or 2.5 million KRW $<$ property damage cost ≤ 1 billion KRW). Level C stands for evident injury ($1 <$ injured persons ≤ 5 or 300 thousand won $<$ property damage cost ≤ 2.5 million KRW), and Level D denotes property damage only (PDO) (damage cost ≤ 300 thousand KRW) [35].

In total, 107,173 observations representing crashes that occurred from 2008 to 2017 on the expressways were obtained from KEC. To achieve the aim of our study, we extracted truck-involved crash observations from the database containing all crashes on the expressway. The number of truck-involved crash observations spanning all 38 expressway routes within the study period in South Korea was 19,038. We arranged the causes of crashes into groups such as vehicle faults (tire puncture) and driver's faults (negligence, overspeeding, and drowsy driving). Table 1 shows the list of truck-involved crash information and summarizes the frequency distribution of truck crash incidents.

The crash database to be used for the ARM analysis contains 17 explanatory items with 98 subitems. The table shows that truck frequency is lowest on Sundays but is almost similar throughout the other days of the week. Also, truck-involved crashes are persistent during the day (6 AM to 11:59 PM: 30.3%; 12 PM to 5:59 PM: 35.2%). Considering the variable for the months, it shows that the majority of the crashes occurred during summer (June, July, and August). In this study, horizontal alignment is grouped based on their lengths. Horizontal alignment is termed straight when there is no curve. Road surfaces which are horizontally aligned to the left or right side are termed left or right curves, respectively. Their curve length, which ranges from 500 m to 1000 m, is also indicated. Vertical alignments are grouped into upward slopes (crest curves) and downward slopes (sag curves), and their slopes or grades are indicated as shown in Table 1. Most of the crashes occurred on the mainline (64.4%) and straight and flat roads (horizontal alignment, straight: 75.9%; vertical alignment, no slope: 63.2%). Approximately 69.3% of all the truck-involved crashes were of severity level D, and most crashes occurred under fine weather conditions (63.2%). In terms of driver-specific

Table 1: Summary of freight-vehicle crashes.

Factor		Frequency	%
Month	January	1,331	7.0
	February	1,187	6.2
	March	1,392	7.3
	April	1,561	8.2
	May	1,656	8.7
	June	1,704	9.0
	July	1,974	10.4
	August	1,877	9.9
	September	1,712	9.0
	October	1,637	8.6
	November	1,477	7.8
	December	1,530	8.0
Time	12 AM–5:59 AM	3,105	16.3
	6 AM–11:59 PM	5,776	30.3
	12 PM–5:59 PM	6,695	35.2
	6 PM–11:59 PM	3,462	18.2
Route	Gyeongbu-line	3,470	18.2
	Sehaean-line	1,742	9.2
	Youngdong-line	1,599	8.4
	Namhae-line	1,501	7.9
	Jungbunaeruk-line	1,403	7.4
	Jungbu-line	1,394	7.3
	Jungang-line	1,238	6.5
	Others	6,691	35.1
Segment	Mainline	12,256	64.4
	Toll gate	3,113	16.4
	Ramp	2,548	13.4
	Tunnel	744	3.9
	Rest area	273	1.4
	Others	104	0.5
Severity	A	36	0.2
	B	808	4.2
	C	5,000	26.3
	D	13,194	69.3
Cause	Negligence	6,051	31.8
	Overspeeding	3,683	19.3
	Drowsy	3,039	16.0
	Tire puncture	1,353	7.1
	Falling detritus	1,140	6.0
	Unsafe distance	591	3.1
	Improper passing	241	1.3
	Others	2,940	15.4
Week of day	Sunday	1,481	7.8
	Monday	3,027	15.9
	Tuesday	3,177	16.7
	Wednesday	2,940	15.4
	Thursday	2,943	15.5
	Friday	3,040	16.0
	Saturday	2,430	12.8
Weather	Fine	12,040	63.2
	Rainy	3,516	18.5
	Cloudy	2,800	14.7
	Snowy	595	3.1
	Others	87	0.5
Crash types	Vehicle-facility	11,619	61.0
	Vehicle-vehicle	3,638	19.1
	Type-others	3,617	19.0
	Vehicle-pedestrian	164	0.9
Number of vehicles	Single-vehicle-involved	13,327	70.0
	Multivehicle-involved	5,711	30.0
[1]Horizontal alignment	Straight	14,450	75.9
	RCL > 1,000 m	2,324	12.2
	LCL > 1,000 m	2,155	11.3
	RCL < 500 m	46	0.2
	500m ≤ RCL ≤ 1,000 m	25	0.1
	500m ≤ LCL ≤ 1,000 m	20	0.1
	LCL < 500 m	18	0.1
[2]Vertical alignment	No slope	12,038	63.2
	1% ≤ DS ≤ 3%	1,957	10.3
	1% ≤ US ≤ 3%	1,704	9.0
	DS < 1%	1,048	5.5
	US < 1%	968	5.1
	DS > 3%	689	3.6
	US > 3%	634	3.3
Median	Fixed wall (127 cm)	6,459	33.9
	Fixed wall (81 cm)	3,953	20.8
	Guardrail	1,624	8.5
	Landscape	442	2.3
	No median	4,141	21.8
	Others	5,419	12.7
Shoulder	Guardrail	8,057	42.3
	Concrete guard	1,783	9.4
	Guard fence	129	0.7
	Guard pipe	39	0.2
	No guardrail	5,639	29.6
	Others	3,391	17.8
Vehicle weights (ton)	>3.5 t and ≤8.5 t	7,834	41.1
	≤3.5 t	6,130	32.2
	>8.5 t	5,074	26.7
Vehicle types	Cargo truck	11,241	59.0
	Box truck	3,876	20.4
	Logging truck	2,950	15.5
	Tanker	971	5.1
Driver's age group	20s	4,315	22.7
	30s	2,540	13.3
	40s	4,841	25.4
	50s	5,369	28.2
	Over 60s	1,973	10.4

[1]RCL and LCL indicate right curve length and left curve length, respectively.
[2]DS and US indicate downward slope and upward slope, respectively.

variables, older drivers were observed to have more crashes (age group in the 50s: 28.2%), and drivers' negligence was found to be the majority cause of truck-involved crashes in Korea. Cargo trucks and heavy trucks were observed to be among the trucks highly prone to crashes.

4. Methodology

Association rules mining (ARM) is a data mining technique that involves identifying a set of items that occur together in an event [33]. In terms of truck safety, it is seen as a technique employed to determine groups of crash characteristics that are observed in a truck crash. The Apriori

algorithm is well known for discovering association rules due to its exploratory and easy to understand nature [5, 9]. The details of the algorithm are given as follows.

Assuming $I = \{i_1, i_2, \ldots, i_n\}$ is a set of n crash attributes called items (a set of crash characteristics for each truck crash record) and $D = \{t_1, t_2, \ldots, t_m\}$ is a database of truck crash information, such that each crash information in D has a unique identification ID and each $t_i \in I$ represents each truck crash record made up of a subset of items selected from I. Following Agrawal et al. [7], we define a rule as an implication of the form $A \Longrightarrow B$ (for example, {Driver's age group: 20s, Weather: Rainy} $\Longrightarrow$ {Crash type: Vehicle-vehicle}), where A (Driver's age group: 20s, Weather: Rainy) and B (Crash type: Vehicle-vehicle) are itemsets which belong to D, A is the antecedent on the left hand side (LHS), and B is the consequent on the right hand side (RHS) of the rule, and $A \subset I$, $B \subset I$, and $A \cap B = \{\}$.

Finding association rules using the Apriori algorithm involves a "bottom-up" approach. The Apriori algorithm employs the philosophy that a k-itemset is frequent if and only if each item in the itemset is also frequent [7]. Two main steps are involved when mining interesting rules from a dataset. The first step involves frequent itemset generation. The algorithm first scans the database to identify itemsets which satisfy some predefined minimum *support*. In the next stage, the algorithm generates rules above a predetermined minimum *confidence*.

The *support* and *confidence* are essential concepts in ARM used for selecting important rules from all possible rules. The *support* shows how frequently a combination of antecedent and consequent of a rule occurs together in the database, and the *confidence* measures the credibility or the strength of the rule by estimating the probability $P(A \mid B)$, interpreted as the share of cases in which the consequent occurs given that the antecedent has occurred [6, 8]. The *support* and *confidence* can be estimated using the following equations:

$$support\,(A \Longrightarrow B) = P(A \cap B) = \frac{|A \cup B|}{|D|}, \tag{1}$$

$$confidence\,(A \Longrightarrow B) = \frac{support\,(A \Longrightarrow B)}{support\,(A)} = \frac{P(A \cap B)}{P(A)}. \tag{2}$$

It is essential to know that, in a single rule, there could be multiple itemsets as either antecedents and or consequents. The association rule $A \Longrightarrow B$ should satisfy predefined minimum thresholds α and β such that $support\,(A \Longrightarrow B) \geq \alpha$ and $confidence\,(A \Longrightarrow B) \geq \beta$. They are to be adjusted until interesting rules are observed [36]. If a rule $A \Longrightarrow B$ satisfies the minimum support condition with *support* value s, then it can be interpreted as s% of the crash records in the database D containing $A \cup B$. Also, if a rule $A \Longrightarrow B$ holds with *confidence* c, then c% of the crash records in the database D that contain A also contain B [37].

Depending on the dataset being studied, ARM algorithms may produce a large set of rules that satisfies the predefined thresholds for both α and β [6]. Lee et al. [5] argue that *confidence* fails to take the baseline frequency of the consequent into consideration, rendering it deficient. As such, another measure known as *lift* was proposed to overcome the aforementioned limitations by including the frequency of the consequent in its equation as in the following formula:

$$lift\,(A \Longrightarrow B) = \frac{confidence\,(A \Longrightarrow B)}{support\,(A)} = \frac{P(A \cap B)}{P(A)P(B)}. \tag{3}$$

The *lift* of the rule $A \Longrightarrow B$ shows how much the probability of B will increase if A occurs [5]. There are three instances. When $lift\,(A \Longrightarrow B) > 1$, then there exists a positive interdependence between the antecedent and consequent and the rule is seen as valuable. When $lift\,(A \Longrightarrow B) < 1$, then there is a negative interdependence between the antecedent and the consequent. Finally, when $lift\,(A \Longrightarrow B) = 1$, then A and B are independent, and there is no correlation between them. The higher the *lift* measure, the higher the interestingness of the generated rules. With the aid of this measure, we sorted the rules that met the minimum *support* and *confidence* thresholds.

The Apriori algorithm is explained in the flowchart illustrated in Figure 1, and the steps taken in generating interesting rules are summarized as follows:

(1) Initially, scan database and find all frequent items.

(2) Generate *support* for the items. Items are discarded if they do not meet the minimum *support* thresholds (α).

(3) The remaining items that met the predetermined *support* are used to generate all possible itemset configurations.

(4) From the frequent itemsets, find temporal association rules that satisfy the predetermined minimum *confidence* (β).

(5) Generate *lift* for the frequent itemsets. Items with lift greater that 1 are selected as strong association rules.

5. Results and Discussion

In this study, we employed the Apriori algorithm to generate association rules from the characteristics and crash contributory factors in the dataset. To obtain interesting rules, determining optimum support α and confidence β thresholds is very crucial. Thus, we conducted several trials using various combinations of α and β. As shown in Table 2, it is observed that the number of rules decreases as the thresholds increase. For $\alpha = 1\%$ and $\beta = 10\%$, a total of 851,955 rules were generated. On the other hand, when $\alpha = 14\%$ and $\beta = 100\%$, only 1 rule was obtained. Clearly, setting very low threshold values will yield a huge number of uninteresting rules.

In traffic crash analysis, several researchers used minimum support threshold values in the range of 1–4% and minimum confidence threshold values in the range of 10–70% [6, 9, 36]. In our study, we observed an interesting relationship between the number of rules produced and α and β values. From the graphs in Figures 2(a) and 2(b), there is a general sharp decline in the number of rules generated when both α and β are very negligible. However, there is a

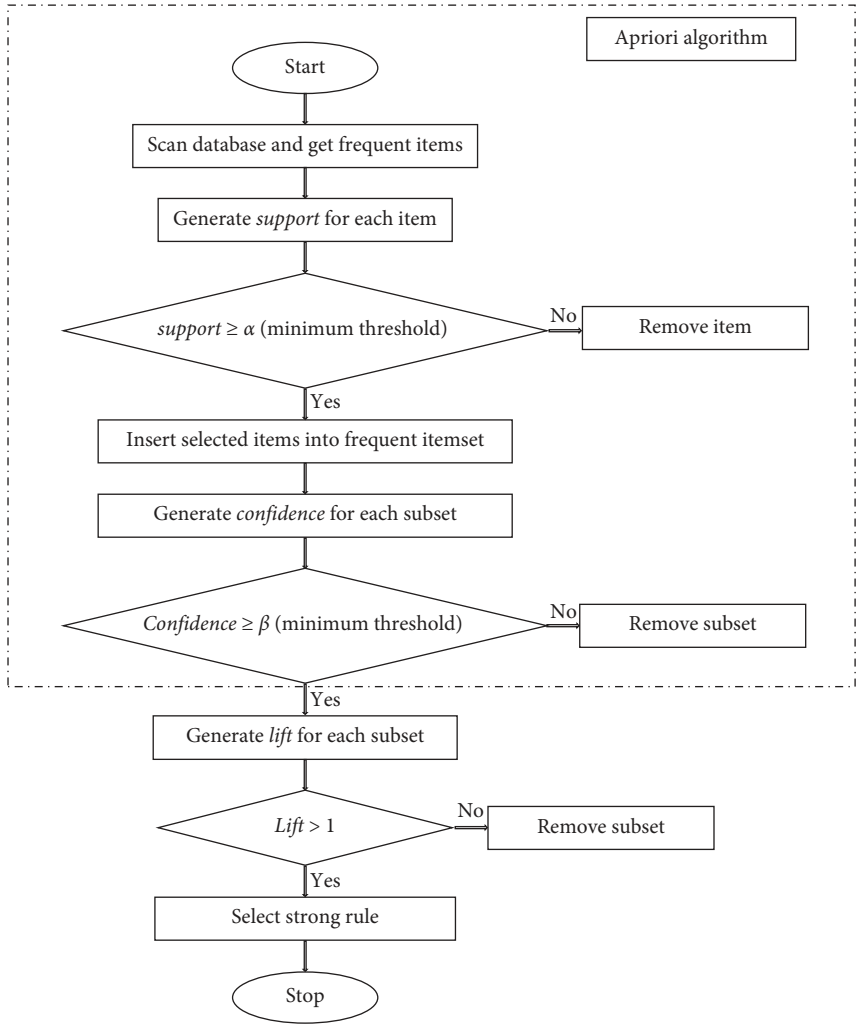

Figure 1: Flowchart of the association rule mining process.

Table 2: Number of rules by combination of minimum support and confidence.

α/β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.01	851,955	783,385	705,858	641,263	570,749	473,965	340,999	206,040	102,130	6,471
0.02	215,951	202,489	183,575	167,520	149,725	**90,951**	86,910	51,382	24,605	803
0.03	90,379	85,715	78,150	71,446	63,969	53,177	36,305	21,123	9,674	228
0.04	46,854	44,698	40,979	37,524	33,692	27,962	18,619	10,525	4,421	97
0.05	27,952	26,847	24,753	22,690	20,474	17,096	11,187	6,237	2,484	56
0.06	13,775	13,182	12,242	10,934	9,815	8,101	5,417	3,366	1,462	25
0.07	9,183	8,840	8,232	7,334	6,639	5,529	3,641	2,220	835	16
0.08	6,475	6,263	5,827	5,184	4,718	3,932	2,561	1,525	512	9
0.09	4,741	4,600	4,301	3,835	3,507	2,929	1,873	1,110	358	7
0.1	3,626	3,531	3,312	2,941	2,699	2,267	1,458	872	285	4
0.11	2,763	2,712	2,547	2,266	2,076	1,750	1,115	662	205	1
0.12	2,153	2,122	2,002	1,766	1,629	1,371	858	494	145	1
0.13	1,688	1,667	1,575	1,396	1,284	1,082	656	366	95	1
0.14	1,306	1,288	1,214	1,089	1,001	842	501	273	64	1
0.15	1,073	1,059	1,003	905	831	698	407	207	43	—
0.16	884	876	835	756	696	576	326	160	32	—
0.17	720	713	688	626	578	484	272	135	24	—
0.18	587	582	566	510	479	394	208	95	16	—
0.19	507	505	492	447	421	351	181	84	13	—

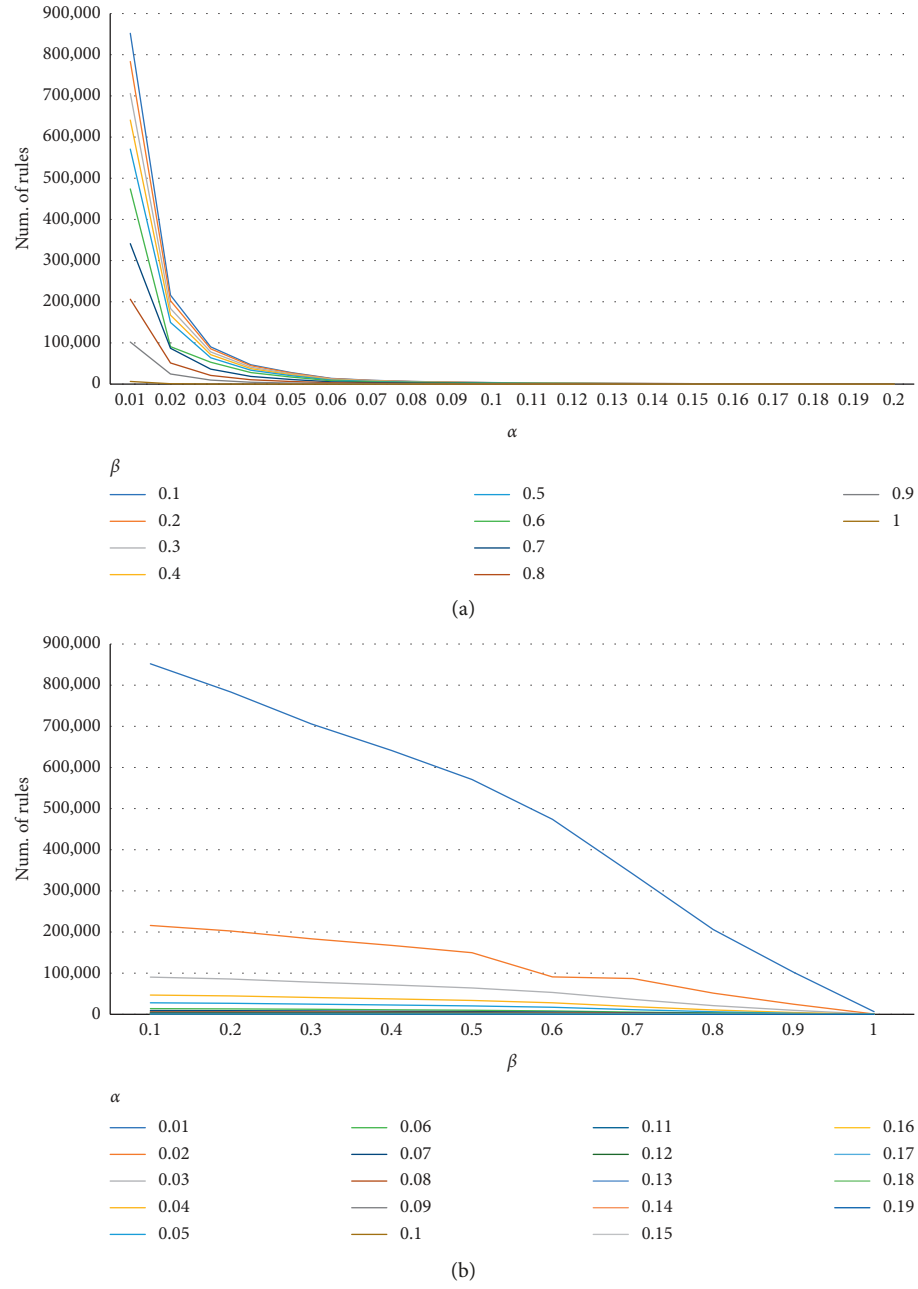

Figure 2: Distribution of association rules based on (a) minimum support (α) and (b) confidence (β) values.

sudden change within the region of $\alpha = 2\%$ and $\beta = 60\%$ in Figures 2(a) and 2(b), respectively. At this point, the number of rules generated is reduced to 90,951.

Since the number of truck-involved crash observations used for this analysis is 19,038, the selected minimum support threshold value (2%) means that, for a crash factor, or a set of crash factors to be considered, it should appear in at least 381 truck crash records. On the other hand, the selected minimum confidence value (60%) means that a generated rule is deemed as credible if it occurs at least 60 percent of the time. Even though the selection of these threshold values is subjective, we are sure that the rules obtained are stable, interesting, and noteworthy.

Statistics of average support, confidence, and lift values for the generated rules given different sizes of itemsets are shown in Table 3 and visualized in Figure 3. The number of rules generated increases with increasing itemset size, and the difference between the total number of rules generated and the number of rules greater than 1 is marginal. For itemsets of size $n = 2$, only 583 rules were generated, of which 517 were greater than 1, and for itemsets of size $n = 3$, 6,891 rules were generated with 6,323 rules having significant interdependence between the antecedents and consequents. From the table, the maximum number of items in an itemset was $n = 10$. The total number of rules was 90,951, out of which 88,018 rules had a lift greater than 1. This

Table 3: Average support, confidence, and lift values by subset sizes.

Size of subsets (n)	Number of rules	Number of rules lift >1	Average support values	Average confidence values	Average lift values
2	583	517	0.117	0.696	1.383
3	6,891	6,323	0.060	0.726	1.316
4	27,320	25,704	0.045	0.752	1.379
5	55,812	53,338	0.040	0.770	1.445
6	77,064	74,225	0.037	0.782	1.494
7	86,968	84,044	0.036	0.789	1.526
8	90,159	87,227	0.036	0.792	1.543
9	90,868	87,935	0.036	0.792	1.548
10	90,951	88,018	0.036	0.792	1.549

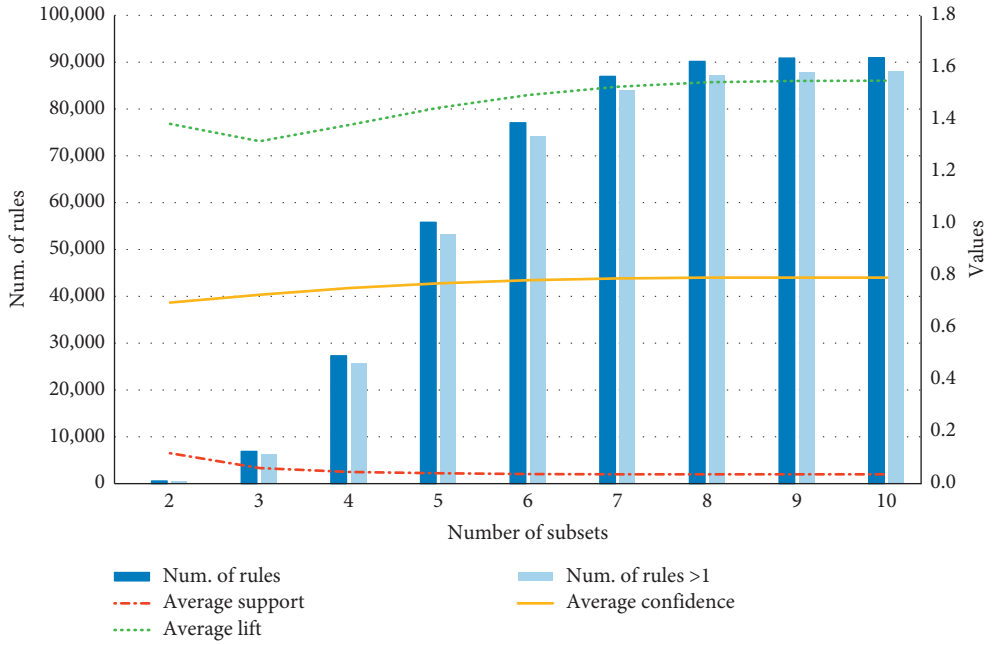

Figure 3: Graph for the association rules mining.

observation had the highest average lift value. According to Weng et al. [6], the higher the lift value, the higher the association between the antecedent and consequent. As the average lift value increases with increasing itemset size, we can say that the interestingness of the association rules increases with increasing itemset size.

5.1. Key Contributory Patterns for Truck-Involved Crashes. In the real world, many factors contribute to each distinct crash. Hence, it is likely to have several items in either the antecedent or consequent. Figure 4 visualizes the frequency of items generated by ARM. Overall, the top ten frequent items in the truck-involved crash database are {Median: Guardrail}, {Horizontal Alignment: Straight}, {Number of vehicles involved: Single vehicle-involved}, {Segment: Mainline}, {Weather: Fine}, {Vertical Alignment: No slope}, {Crash type: vehicle-facility}, {Vehicle type: Cargo truck}, {Median = Fixed wall}, and {Cause: Negligence}, in that order.

The graph-based visualization provided in Figure 5 helps us understand the patterns of the rules generated by ARM. A vertex represents each item, and connections between every pair of vertices represent the relationship between the antecedent and consequent. The link begins from the antecedent and ends on the consequent. From the graph, items with a larger circle have more association compared to the items with a smaller circle. The results show that driving cargo trucks on a straight and flat mainline section on a fine weather is likely to result in a single-vehicle (SV) crash where the truck runs into a fixed wall or any roadway facility (Figures 4 and 5)

To concentrate on attaining meaningful analysis, we identify critical crash contributory factors that lead to truck-involved crashes at various segment types (mainline, ramp, and toll gate), under different weather conditions (fine and rainy), considering the drivers age (20s and 30s), crash type (vehicle to vehicle and vehicle to facility), driver's faults (negligence and overspeeding), vehicle weight (<3.5 t, 3.5 t and<8.5 t, >8.5 t), and roadway geometry type (straight and flat surfaces). These key variables were those that obtained lift values greater than 1. For this study, we screened out the top 5 rules for each subitem mentioned above and sorted them by descending lift values.

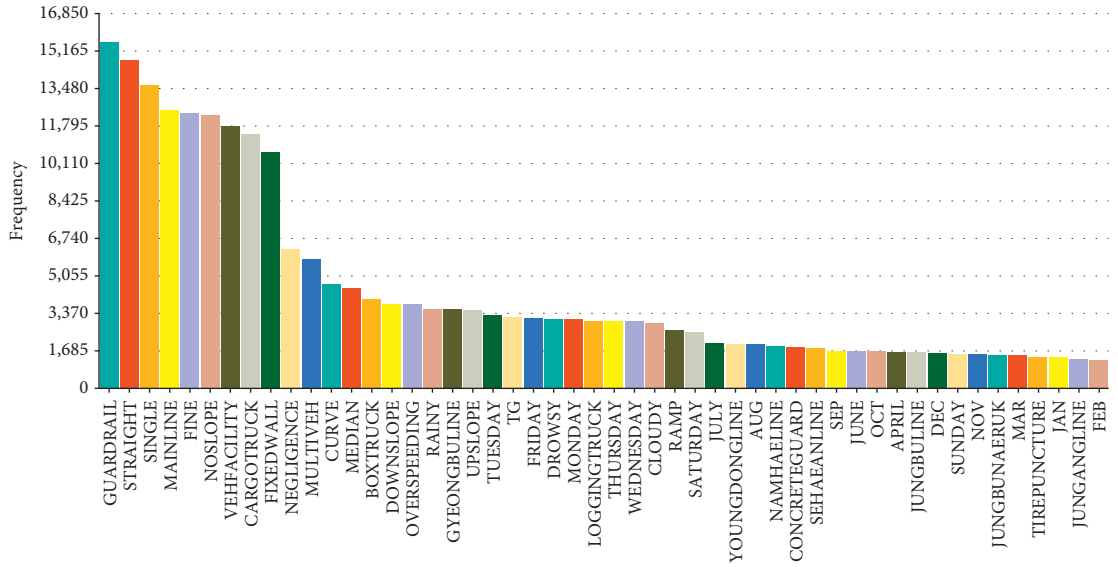

Figure 4: Distribution of frequent keywords.

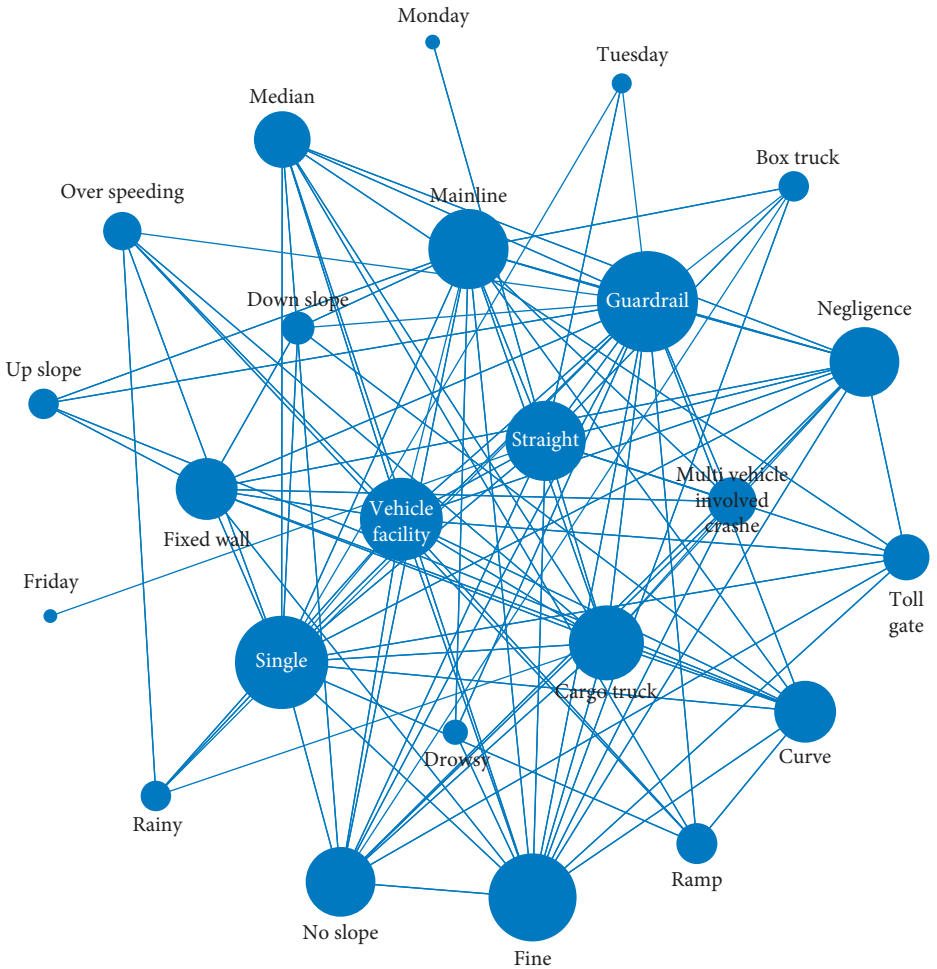

Figure 5: Network graph for association among keywords.

Considering the association rules by the expressway segment type shown in Table 4, it is easy to see that the antecedents (LHS) of the high-lift rules relating to median type, roadway geometry, and the number of vehicles involved in a crash: {Median: fixed wall (127 cm)}, {Vertical alignment: no slope}, {Horizontal alignment: straight},

Table 4: Association rules by segment types.

LHS	RHS	Itemset	α	β	Lift
{Median: fixed wall (127 cm), Region: Daegu, Route: Gyeongbu-line, Vertical alignment: no slope}	{Segment: mainline}	5-itemset	0.0200	1.0000	1.6127
{Horizontal alignment: straight, Median: fixed wall (127 cm), Number of vehicles: multi-vehicle involved, Route: Gyeongbu-line, Vertical alignment: no slope}		6-itemset	0.0261	0.9980	1.6094
{Level: D, Median: fixed wall (81 cm), Number of vehicles: multi-vehicle involved, Shoulder: guardrail}		5-itemset	0.0228	0.9977	1.6090
{Horizontal alignment: straight, Median: fixed wall (127 cm), Number of vehicles: multi-vehicle involved, Route: Gyeongbu-line, Weather: fine}		6-itemset	0.0218	0.9976	1.6088
{Driver age: 20s, Median: fixed wall (127 cm), Number of vehicles: multi-vehicle involved, Vehicle type: cargo truck}		5-itemset	0.0207	0.9975	1.6086
{Cause: over-speeding, Crash type: vehicle-facility, Median: no median, Number of vehicles: single vehicle, Vehicle type: cargo truck}	{Segment: ramp}	6-itemset	0.0215	0.7518	5.7603
{Cause: over-speeding, Crash type: vehicle -facility, Level: D, Median: no median, Number of vehicles: single vehicle}		6-itemset	0.0242	0.7508	5.7524
{Cause: over-speeding, Level: D, Median: no median, Number of vehicles: single vehicle}		5-itemset	0.0262	0.7500	5.7462
{Cause: over-speeding, Crash type: vehicle -facility, Median: no median, Number of vehicles: single vehicle}		5-itemset	0.0294	0.7477	5.7283
{Cause: over-speeding, Level: D, Median: no median, Number of vehicles: single vehicle, Shoulder: no guardrail}		6-itemset	0.0214	0.7459	5.7147
{Cause: negligence, Crash type: vehicle -facility, Level: D, Median: others, Number of vehicles: single vehicle, Shoulder: others, Vertical alignment: no slope}	{Segment: TG}	8-itemset	0.0202	0.9413	5.9979
{Median: others, Number of vehicles: single vehicle, Shoulder: others, Vehicle weight: >8.5 t, Vertical alignment: no slope}		6-itemset	0.0223	0.9361	5.9648
{Crash type: vehicle -facility, Median: others, Shoulder: others, Vehicle weight: >8.5 t}		5-itemset	0.0201	0.9340	5.9512
{Cause: negligence, Crash type: vehicle -facility, Median: others, Number of vehicles: single vehicle, Shoulder: others, Vertical alignment: no slope}		7-itemset	0.0234	0.9329	5.9444
{Driver age: 50s, Median: others, Number of vehicles: single vehicle, Shoulder: others, Vertical alignment: no slope}		6-itemset	0.0218	0.9103	5.8002

{Median: fixed wall (81 cm)}, and {Number of vehicles: multi-vehicle involved} are highly associated with truck crashes that occur on the mainline. At the ramp section of the expressway, the most frequent items associated with crashes include {Cause: overspeeding}, {Crash type: vehicle-facility}, {Median: no median}, and {Number of vehicles involved: single}. These crashes are mostly of severity level D, which is consistent with findings in the literature [19]. At the toll gate section, the items mostly associated with truck crashes include {Crash type: vehicle-facility}, {Number of vehicles: single vehicle}, {Vertical alignment: no slope}, and {Vehicle weight: >8.5 t}. The results show that heavy trucks are more likely to crash into facilities around the toll booth areas with flat (no slope) road surfaces. These crashes mostly result in SV crashes with a severity level of D. In Table 4, the rule with the highest lift value (5.99) is {Cause: negligence, Crash type: vehicle-facility, Level: D, Median: others, Number of vehicles: single vehicle, Shoulder: others, Vertical alignment: no slope} $\Longrightarrow$ {Segment: TG}. This rule indicates that if a single vehicle-facility low severity crash happened on a roadway with no slope as a result of the driver's negligence, it is more likely to have occurred at the toll gate section of the expressway.

Other interesting rules concerning the weather condition are exhibited in Table 5. The analysis uncovers that drowsy driving and driving from 12 noon to 5 PM are the main factors associated with fine weather crashes. Also, items highly associated with truck crashes during the rainy weather are {Cause: over-speeding} and {Vehicle weight: >3.5 t and<8.5 t}. Considering the weather condition, the rule with the highest lift (4.2256) {Cause: over-speeding, Crash type: vehicle-facility, Number of vehicles: single, Segment: mainline, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck} $\Longrightarrow$ {Weather: rainy} shows that an SV-

Table 5: Association rules by weather condition.

LHS	RHS	Itemset	α	β	Lift
{Cause: drowsy, Segment: mainline, Time: *t*12–17}	{Weather: fine}	4-itemset	0.0258	0.8466	1.3941
{Cause: drowsy, Horizontal alignment: straight, Time: *t*12–17}		4-itemset	0.0235	0.8312	1.3688
{Cause: poor loading}		2-itemset	0.0214	0.8124	1.3378
{Crash type: vehicle- vehicle, Horizontal alignment: straight, Number of vehicles: multi-vehicle involved, Time: *t*12–17, Vertical alignment: no slope}		6-itemset	0.0226	0.8117	1.3367
{Cause: drowsy, Driver age: 20s}		3-itemset	0.0241	0.8024	1.3215
{Cause: over-speeding, Month: JULY}	{Weather: rainy}	3-itemset	0.0213	0.7879	4.3533
{Cause: over-speeding, Crash type: vehicle-facility, Number of vehicles: single, Segment: mainline, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck}		6-itemset	0.0210	0.7648	4.2256
{Cause: over-speeding, Crash type: vehicle -facility, Number of vehicles: single, Segment: mainline, Vehicle weight: >3.5 t and<8.5 t}		6-itemset	0.0281	0.7535	4.1632
{Cause: over-speeding, Crash type: vehicle -facility, Segment: mainline, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck}		6-itemset	0.0230	0.7526	4.1580
{Cause: over-speeding, Crash type: vehicle -facility, Horizontal alignment: straight, Number of vehicles: single, Vehicle weight: >3.5 t and<8.5 t}		6-itemset	0.0246	0.7301	4.0338

facility crash involving a medium-weight cargo truck that occurred on the mainline due to over speeding is likely to occur on a rainy day.

The reason for this association could be that roads are slippery on rainy days. Hence, overspeeding makes it difficult for drivers to control trucks in case of an emergency. As the mainline segment of expressways is always busy, drivers of heavy trucks who overspeed on rainy days are highly likely to have single-vehicle crashes. This observation is consistent with previous research on truck safety [18, 35].

Rules with {Driver age: 20s & 30s} on the RHS were the only ones with a lift value greater than 1. Again, we selected the top 5 of the rules and arranged them in descending order, as presented in Table 6. The frequent items from all the rules generated in this category are {Horizontal alignment: straight}, {Region: Changwon}, and {Vertical alignment: no slope}. The rule with the highest lift value (3.6399) is {Horizontal alignment: straight, Region: Changwon, Route: Namhae-line, Vertical alignment: no slope} $\Longrightarrow$ {Driver age: 20s & 30s}. This rule can be explained as 2.06% of the truck-involved crashes that occurred on leveled and straight expressway sections in the Changwon region of South Korea and on the Namhae-line route are highly associated with drivers in their 20s and 30s. Also, out of all the crashes in the dataset which have crash records of the itemset {Horizontal alignment: straight, Region: Changwon, Route: Namhae-line, Vertical alignment: no slope}, 79.84% of the drivers were between the ages of 20 and 30.

These results present interesting findings. It is worth noting that Changwon city is home to the Changwon Industrial Complex, which comprises a hub of heavy industrial factories such as LG Electronics and GM Korea and employs almost 100,000 people of which the majority are young. As such, it is reasonable to infer that the young drivers in that region are mostly involved in truck crashes since they form the majority of the city's dwellers. The results also showed that young drivers are associated with truck crashes on straight and leveled roads, especially when the weather is good. The findings are consistent with research conducted by Yu et al. [31], which mentions that drivers tend to indulge in risky driving when the weather and road conditions are right. The rules discussed above are shown in Table 6.

Table 6: Association rules by driver's age group.

LHS	RHS	Itemset	α	β	Lift
{Horizontal alignment: straight, Region: Changwon, Route: Namhae-line, Vertical alignment: no slope}	{Driver age: 20s & 30s}	4-itemset	0.0206	0.7984	3.6399
{Horizontal alignment: straight, Region: Changwon, Weather: fine}		4-itemset	0.0212	0.7860	3.5835
{Horizontal alignment: straight, Region: Changwon, Vertical alignment: no slope}		4-itemset	0.0246	0.7804	3.5578
{Region: Changwon, Vertical alignment: no slope}		4-itemset	0.0247	0.7213	3.2885
{Region: Changwon, Vehicle type: cargo truck}		4-itemset	0.0206	0.6907	3.1489

Table 7 lists the top 5 rules that contain highly associated crash characteristics for both vehicle-vehicle crash type and

Table 7: Association rules by crash type.

LHS	RHS	Itemset	α	β	Lift
{Level: B, Number of vehicles: multi-vehicle involved, Segment: mainline}	{Crash type: vehicle-vehicle}	4-itemset	0.0232	0.8384	4.5935
{Cause: unsafe distance, Number of vehicles: multi-vehicle involved}		3-itemset	0.0216	0.8175	4.4787
{Cause: drowsy, Median: fixed wall (127 cm), Number of vehicles: multi-vehicle involved, Segment: mainline, Weather: fine}		6-itemset	0.0216	0.8175	4.4787
{Cause: drowsy, Horizontal alignment: straight, Median: fixed wall (127 cm), Number of vehicles: multi-vehicle involved, Segment: mainline}		6-itemset	0.0204	0.7918	4.3384
{Cause: drowsy, Horizontal alignment: straight, Number of vehicles: multi-vehicle involved, Segment: mainline, Vertical alignment: no slope}		6-itemset	0.0222	0.7645	4.1885
{Cause: negligence, Driver age: 50s, Horizontal alignment: straight, Level: D, Number of vehicles: single, Segment: TG}	{Crash type: vehicle-facility}	7-itemset	0.0209	0.9900	1.6794
{Cause: negligence, Driver age: 50s, Horizontal alignment: straight, Level: D, Number of vehicles: single, Segment: TG, Vertical alignment: no slope}		8-itemset	0.0207	0.9900	1.6793
{Driver age: 50s, Horizontal alignment: straight, Number of vehicles: single, Segment: TG, Vehicle type: cargo truck}		6-itemset	0.0210	0.9852	1.6711
{Driver age: 50s, Horizontal alignment: straight, Number of vehicles: single, Segment: TG, Vehicle type: cargo truck, Vertical alignment: no slope}		6-itemset	0.0209	0.9851	1.6710
{Cause: negligence, Median: no median, Number of vehicles: single, Segment: TG, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck}		7-itemset	0.0208	0.9851	1.6710

vehicle-facility crash type. In the first category with {Crash type: vehicle-vehicle} on the RHS, we observe that the most common antecedents on the LHS are {Segment: mainline}, {Number of vehicles: multi-vehicle involved}, and {Cause: drowsy}. For rules with {Crash type: vehicle-facility} on the RHS, we observe items such as {Segment: TG}, {Horizontal alignment: straight}, {Driver age: 50s}, {Number of vehicles: single}, and {Cause: negligence}. The frequent occurrence of these antecedents shows that drowsy driving is the leading cause of vehicle-vehicle crashes, whereas negligence is the main cause of vehicle-facility crashes. Vehicle-vehicle crashes are more prone to occur on the mainline, whereas vehicle-facility crashes are more probable to occur around the toll gate section of the expressway. The highest lift value rules in both categories ({Level: B, Number of vehicles: multi-vehicle involved, Segment: mainline} $\Longrightarrow$ {Crash type: vehicle-vehicle}; {Cause: negligence, Driver age: 50s, Horizontal alignment: straight, Level: D, Number of vehicles: single, Segment: TG} $\Longrightarrow$ {Crash type: vehicle-facility}). From these results, decision-makers must make policies directed at checking driver attitudes when driving and during work hours of truck drivers.

Another set of interesting rules related to driver's faults are presented in Table 8. Two sets of consequents on the RHS were the only factors that had lift values greater than 1. We selected the top 5 rules under each category, as displayed in the table. From the rules with RHS {Cause: negligence}, the common items on the LHS are {Segment: TG}, {Vertical alignment: no slope}, {Horizontal alignment: straight}, and {Crash type: vehicle-facility}. The study results show that crashes that occurred at the toll gate sections that have straight and leveled surfaces are often as a result of driver's negligence. As accounted by previous studies, the tendency of indulging in risky driving increases when driving on such roads. This observation is because drivers have a good field of view, and the probability of becoming negligent then rises, causing crashes. As shown in Table 8, the crashes which occur during rainy weather are highly likely to be caused by overspeeding, and most of these crashes result in severity level D. In contrast with other studies focused on truck-involved crashes [15], our study identified that crashes that occurred during rainy weather and resulted in severity level D are associated with overspeeding. It reminds us of the need to enforce speed regulations and use speed control features as it is cited as being highly efficient in decreasing the frequency and severity of crashes [23].

Table 9 shows the association rules obtained under the vehicle weight category. Under the category, the rule with the highest lift has {Vehicle weight: >8.5 t} as its consequent. Its antecedent {Vehicle type: logging truck, Weather: cloudy} shows that logging trucks that are involved in crashes are likely to be heavy vehicles, and SV cargo truck crashes that occur at the ramp section resulting in severity level D are likely to have lightweight trucks involved (RHS is {Vehicle weight: <3.5 t}). Also, cargo truck crashes on leveled roads that result in severity

Table 8: Association rules by driver's faults.

LHS	RHS	Itemset	α	β	Lift
{Crash type: vehicle-facility, Horizontal alignment: straight, Median: no median, Number of vehicles: single, Segment: TG, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck, Vertical alignment: no slope}	{Cause: negligence}	9-itemset	0.0205	0.7377	2.4556
{Level: D, Median: no median, Segment: TG, Vehicle weight: >3.5 t and<8.5 t, Vehicle type: cargo truck, Vertical alignment: no slope}		7-itemset	0.0200	0.7257	2.4155
{Crash type: vehicle-facility, Driver age: 20s, Horizontal alignment: straight, Number of vehicles: single, Segment: TG, Vertical alignment: no slope}		7-itemset	0.0201	0.7254	2.4144
{Crash type: vehicle-facility, Driver age: 20s, Horizontal alignment: straight, Segment: TG, Vertical alignment: no slope}		6-itemset	0.0205	0.7196	2.3950
{Crash type: vehicle-facility, Horizontal alignment: straight, Level: D, Segment: TG, Time: t12–17, Vertical alignment: no slope}		7-itemset	0.0256	0.7001	2.3304
{Crash type: vehicle-facility, Number of vehicles: single, Segment: ramp, Vehicle weight: <3.5 t, Weather: rainy}	{Cause: over-speeding}	6-itemset	0.0202	0.8384	4.4023
{Crash type: vehicle-facility, Level: D, Number of vehicles: single, Segment: ramp, Vehicle type: cargo truck, Weather: rainy}		7-itemset	0.0227	0.8232	4.3223
{Crash type: vehicle-facility, Number of vehicles: single, Segment: ramp, Vertical alignment: no slope, Weather: rainy}		5-itemset	0.0215	0.8131	4.2694
{Crash type: vehicle-facility, Level: D, Time: t6–11, Vehicle type: cargo truck, Weather: rainy}		6-itemset	0.0202	0.7237	3.7998
{Crash type: vehicle-facility, Number of vehicles: single, Segment: mainline, Shoulder: guardrail, Weather: rainy}		6-itemset	0.0283	0.7231	3.7969

level D are expected to be a medium-weight truck (RHS is {Vehicle weight: >3.5 t and <8.5 t}).

The antecedent {Route: Gyeongbu-line} is highly associated with crashes involving heavy trucks. The Gyeongbu line is the longest expressway line in South Korea with a speed limit of 100–110 km/h. Due to its importance, it currently stands as the most used line with an annual average daily traffic (AADT) of 1,335,770 [35]. It also has the highest number of toll booths compared to all the other expressway lines in South Korea. As illustrated in previous rules, the high frequency of antecedents like {Segment: TG}, {Cause: vehicle-facility}, {Cause: negligence}, and {Number of vehicles: single} shows that many of these crashes are likely to be SV crashes at the toll gate section, which is as a result of the features of the Gyeongbu expressway line. According to Hong et al. [35], many SV truck-involved crashes are expected to occur at toll gate sections, reflecting the results of our study. Thus, decision-makers and planners should pay more attention when designing toll gate sections to accommodate the heavy trucks better. Policymakers should make regulations to ensure the safety of heavy trucks on such vital roadways. Also, the results reiterate the reasons why heavy truck drivers should be careful when using toll gate sections of the expressway line.

Expressway design elements related to the horizontal and vertical alignments have shown remarkable findings. From Table 10, the antecedents on the LHS are somewhat similar in both categories on the RHS ({Horizontal alignment: straight} and {Vertical alignment: no slope}). Based on the frequency of the antecedents, we observed that {Segment: TG} and {Median: no median} are highly associated with the consequent {Horizontal alignment: straight}. On the other hand, {Horizontal alignment: straight} and {Median: no median} are the most frequent antecedents associated with the consequent {Vertical alignment: no slope}. It is also significant to note that the items {Vehicle weight: >8.5 t}, {Time: t6–11} and {Segment: TG} are common to both consequents {Horizontal alignment: straight} and {Vertical alignment: no slope}.

The association rules indicate that heavy trucks traveling on the expressway in the morning hours of the day are likely to be involved in vehicle-facility SV crashes at the toll gate section when the roadway surface is straight with no slope. Previous rules depict that drivers are negligent when plying roads with leveled and slopeless surfaces due to their improved field of view. Also, since many trips are made during peak hours, AADT is expected to increase as truck drivers rush to complete their trips. This situation increases their

Table 9: Association rules by vehicle weight.

LHS	RHS	Itemset	α	β	Lift
{Crash type: vehicle-facility, Horizontal alignment: straight, Level: D, Number of vehicles: single, Segment: ramp, Vehicle type: cargo truck, Vertical alignment: no slope}	{Vehicle weight: <3.5 t}	8-itemset	0.0201	0.7114	2.2851
{Crash type: vehicle-facility, Level: D, Number of vehicles: single, Segment: ramp, Vehicle type: cargo truck, Vertical alignment: no slope}		7-itemset	0.0265	0.6852	2.2011
{Number of vehicles: single, Segment: ramp, Vehicle type: cargo truck, Weather: rainy}		5-itemset	0.0228	0.6845	2.1989
{Cause: over-speeding, Level: D, Number of vehicles: single, Segment: ramp, Vehicle type: cargo truck}		6-itemset	0.0269	0.6702	2.1527
{Driver age: 60s, Horizontal alignment: straight, Level: D, Vehicle type: cargo truck}		5-itemset	0.0211	0.6442	2.0694
{Horizontal alignment: straight, Shoulder: guardrail, Vehicle type: cargo truck}	{Vehicle weight: >3.5 t and <8.5 t}	4-itemset	0.0288	0.9649	2.4331
{Horizontal alignment: straight, Level: D, Segment: mainline, Shoulder: guardrail, Vehicle type: cargo truck}		6-itemset	0.0213	0.9621	2.4261
{Crash type: vehicle-facility, Horizontal alignment: straight, Number of vehicles: single, Vehicle type: cargo truck}		5-itemset	0.0226	0.8515	2.1472
{Cause: negligence, Level: D, Median: no median, Segment: TG, Vehicle type: cargo truck, Vertical alignment: no slope}		7-itemset	0.0200	0.7175	1.8094
{Level: D, Median: no median, Segment: TG, Vehicle type: cargo truck, Vertical alignment: no slope, Weather: fine}		7-itemset	0.0202	0.6912	1.7430
{Vehicle type: logging truck, Weather: Cloudy}	{Vehicle weight: >8.5 t}	3-itemset	0.0214	1.0000	3.9199
{Vehicle type: logging truck, Week of day: Thursday}		3-itemset	0.0248	1.0000	3.9199
{Route: Gyeongbu-line, Vehicle type: logging truck}		3-itemset	0.0244	1.0000	3.9199
{Segment: TG, Vehicle type: logging truck}		3-itemset	0.0308	1.0000	3.9199
{Route: Gyeongbu-line, Vehicle type: logging truck}		3-itemset	0.0244	1.0000	3.9199

Table 10: Association rules by geometry.

LHS	RHS	Itemset	α	β	Lift
{Median: no median, Segment: TG, Vehicle weight: >8.5 t}	{Horizontal alignment: straight}	4-itemset	0.0222	1.0000	1.6416
{Median: no median, Segment: TG, Time: t6–11}		4-itemset	0.0243	1.0000	1.6416
{Driver age: 20s, Region: Changwon, Route: Namhae-line, Vertical alignment: no slope}		5-itemset	0.0206	1.0000	1.6416
{Crash type: vehicle-facility, Driver age: 20s, Median: no median, Segment: TG}		5-itemset	0.0202	1.0000	1.6416
{Crash type: vehicle-facility, Median: no median, Segment: TG, Time: t6–11}		4-itemset	0.0227	1.0000	1.6416
{Driver age: 20s, Median: no median, Segment: TG, Shoulder: no guardrail}	{Vertical alignment: no slope}	4-itemset	0.0212	0.9975	1.6574
{Driver age: 20s, Horizontal alignment: straight, Median: no median, Number of vehicles: single, Segment: TG}		6-itemset	0.0210	0.9975	1.6573
{Cause: negligence, Horizontal alignment: straight, Level: D, Segment: TG, Vehicle weight: >8.5 t}		6-itemset	0.0210	0.9975	1.6573
{Crash type: vehicle -facility, Horizontal alignment: straight, Median: no median, Shoulder: no guardrail, Vehicle type: cargo truck, Weather: rainy}		7-itemset	0.0201	0.9871	1.6400
{Cause: negligence, Horizontal alignment: straight, Number of vehicles: single, Segment: TG, Time: t6–11}		6-itemset	0.0241	0.9871	1.6400

chances of crashing into facilities around the toll gate section.

6. Conclusion

Research on truck safety has been extensively conducted, providing policymakers with knowledge of how much crash risk factors affect the severity or the frequency of crashes. However, since each truck-involved crash involves a complex interaction among crash risk factors, traffic safety officers are particularly interested in finding the most influential factors that lead to a crash. Due to the significant impact of truck crashes, it is imperative to find sets of factors that occur together in a single truck crash for use in developing countermeasures to ensure truck safety.

This study bridges the gap in the literature by analyzing patterns of truck crashes and the association between risk factors and crash characteristics from a large database of truck crashes that occurred on the expressways using the ARM technique. Out of a total of 90,951 rules derived through the learning of the crash dataset, 88,018 rules were found to have significant relationships between the crash contributory factors. A summary of the interesting findings from our study is as follows.

First, frequent items in the truck-involved crash database were found to be the guardrail median, straight horizontal alignment, single vehicle-involved, mainline, clear weather, no vertical curve, vehicle-facility collision, cargo truck, and driver's negligence. These factors explain the unique patterns of truck crashes. Segment types considered in this study, namely, mainline, toll gate, and ramp, showed different truck crash characteristics. Truck crashes were found to have mainly occurred on the mainline of the expressway, while crashes on the ramp section were likely to be associated with driver's overspeeding and PDO level. Near toll gates, we found that truck crashes were highly related to trucks of weight greater than 8.5 tons and driver's negligence. Freight vehicle crashes under fine weather conditions showed an association with drowsy driving and driving time between noon and 5 PM, whereas truck crashes on rainy weather were likely to be linked with factors like overspeeding, midsize truck, and single vehicle crash. We also discovered that crashes in a specific region with well-designed roadway geometry, excluding vertical and horizontal curves, were mainly linked with young freight vehicle drivers in their 20s and 30s. Additionally, the result showed the association between young freight vehicle drivers' traffic violations and good conditions of roadways. In terms of crash types, we confirmed that vehicle to vehicle crashes were likely to have the severity of Level B while vehicle to facility crashes are mostly of severity Level D. Besides, the vehicle to vehicle crashes were likely to occur on mainline segments of the expressway, but the vehicle to facility crashes frequently happened near the toll gate entrance.

In this study, we attempted to derive valuable solutions by interpreting the important rules underlying the freight vehicle crash data. As mentioned earlier, ARM is a methodology used to determine the association between freight vehicle crashes and particular risk factors by estimating lift values through the Apriori algorithm. We demonstrated that it is a plausible technique for analyzing patterns and characteristics of freight vehicle crashes. The research results can provide useful insights and suggestions to transport stakeholders, including policymakers and researchers, for creating relevant policies to help reduce freight truck crashes on the expressways.

As a limitation of this study, the raw crash dataset used in this study had no AADT information for each crash observation; therefore, we could not consider the associations between AADT and other crash characteristics and factors. Since AADT is a critical factor to consider in traffic safety analysis, we would consider using it in future studies and comparing the results with findings in the literature.

Data Availability

The crash data used to support the findings of this study have not been made available because the data were supplied by the Korea Expressway Corporation. Requests for access to these data should be made to the Korea Expressway Corporation (KEC).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors wish to thank the Korea Expressway Corporation (KEC) for providing the crash data used in this paper. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (no. 2017R1C1B2010175).

References

[1] B. Naik, L.-W. Tung, S. Zhao, and A. J. Khattak, "Weather impacts on single-vehicle truck crash injury severity," *Journal of Safety Research*, vol. 58, pp. 57–65, 2016.

[2] C. Dong, S. S. Nambisan, S. H. Richards, and Z. Ma, "Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression," *Transportation Research Part A: Policy and Practice*, vol. 75, pp. 30–41, 2015.

[3] A. Behnood and F. L. Mannering, "Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes," *Analytic Methods in Accident Research*, vol. 23, Article ID 100102, , 2019.

[4] G. Tzamalouka, M. Papadakaki, and J. E. Chliaoutakis, "Freight transport and non-driving work duties as predictors of falling asleep at the wheel in urban areas of Crete," *Journal of Safety Research*, vol. 36, no. 1, pp. 75–84, 2005.

[5] S. Lee, Y. Cha, S. Han, and C. Hyun, "Application of association rule mining and social network analysis for understanding causality of construction defects," *Sustainability*, vol. 11, no. 3, Article ID 618, 2019.

[6] J. Weng, J.-Z. Zhu, X. Yan, and Z. Liu, "Investigation of work zone crash casualty patterns using association rules," *Accident Analysis & Prevention*, vol. 92, pp. 43–52, 2016.

[7] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993.

[8] A. Pande and M. Abdel-Aty, "Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool," *Safety Science*, vol. 47, no. 1, pp. 145–154, 2009.

[9] S. Das and X. Sun, "Investigating the pattern of traffic crashes under rainy weather by association rules in data mining," *Transportation Research Board 93rd Annual Meeting*, Transportation Research Board, Washington, DC, USA, 2014.

[10] J. Hong, R. Tamakloe, D. Park, and Y. Choi, "Estimating incident duration considering the unobserved heterogeneity of risk factors for trucks transporting HAZMAT on expressways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 2, pp. 232–242, 2019.

[11] W. R. G. M. Bandara, "Investigation of accidents in Colombo Katunayake expressway," *Master of Engineering in Highway & Traffic Engineering*, University of Moratuwa, Moratuwa, Sri Lanka, 2018.

[12] D. E. Cantor, T. M. Corsi, C. M. Grimm, and K. Özpolat, "A driver focused truck crash prediction model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 46, no. 5, pp. 683–692, 2010.

[13] S. Zhu, P. M. Layde, C. E. Guse et al., "Obesity and risk for death due to motor vehicle crashes," *American Journal of Public Health*, vol. 96, no. 4, pp. 734–739, 2006.

[14] C. Dong, Q. Dong, B. Huang, W. Hu, and S. S. Nambisan, "Estimating factors contributing to frequency and severity of large truck–involved crashes," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 8, Article ID 04017032, 2017.

[15] Z. Zheng, P. Lu, and B. Lantz, "Commercial truck crash injury severity analysis using gradient boosting data mining model," *Journal of Safety Research*, vol. 65, pp. 115–124, 2018.

[16] L.-Y. Chang and F. Mannering, "Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents," *Accident Analysis & Prevention*, vol. 31, no. 5, pp. 579–592, 1999.

[17] J. Pahukula, S. Hernandez, and A. Unnikrishnan, "A time of day analysis of crashes involving large trucks in urban areas," *Accident Analysis & Prevention*, vol. 75, pp. 155–163, 2015.

[18] F. Chen and S. Chen, "Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1677–1688, 2011.

[19] A. Khorashadi, D. Niemeier, V. Shankar, and F. Mannering, "Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis," *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 910–921, 2005.

[20] J. Kim, D. Mehrnaz, and A. Michael, "Analysis of not at-fault truck crashes in Alabama," *International Journal of Traffic and Transportation Engineering*, vol. 6, no. 2, pp. 28–35, 2017.

[21] A. J. Khattak, R. J. Schneider, and F. Targa, "Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions," *TRB 2003 Annual Meeting CD-ROM*, Transportation Research Board, National Research Council, Washington, DC, USA, 2003.

[22] M. M. Ahmed, R. Franke, K. Ksaibati, and D. S. Shinstine, "Effects of truck traffic on crash injury severity on rural highways in Wyoming using Bayesian binary logit models," *Accident Analysis & Prevention*, vol. 117, pp. 106–113, 2018.

[23] Q. Yuan, M. Lu, A. Theofilatos, and Y.-B. Li, "Investigation on occupant injury severity in rear-end crashes involving trucks as the front vehicle in Beijing area, China," *Chinese Journal of Traumatology*, vol. 20, no. 1, pp. 20–26, 2017.

[24] W. Hao, C. Kamga, X. Yang et al., "Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 43, pp. 379–386, 2016.

[25] J. C. Anderson and S. Dong, "Heavy-vehicle driver injury severity analysis by time of week: a mixed logit approach using HSIS crash data," *Institute of Transportation Engineers. ITE Journal*, vol. 87, no. 9, pp. 41–49, 2017.

[26] S.-W. Park and P. P. Jovanis, "Hours of service and truck crash risk," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2194, no. 1, pp. 3–10, 2010.

[27] E. R. Teoh, D. L. Carter, S. Smith, and A. T. McCartt, "Crash risk factors for interstate large trucks in North Carolina," *Journal of Safety Research*, vol. 62, pp. 13–21, 2017.

[28] G. López, J. Abellán, A. Montella, and J. de Oña, "Patterns of single-vehicle crashes on two-lane rural highways in Granada Province, Spain," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2432, no. 1, pp. 133–141, 2014.

[29] J. de Oña, G. López, and J. Abellán, "Extracting decision rules from police accident reports through decision trees," *Accident Analysis & Prevention*, vol. 50, pp. 1151–1160, 2013.

[30] A. T. Kashani, A. Shariat-Mohaymany, and A. Ranjbari, "A data mining approach to identify key factors of traffic injury severity," *Promet-Traffic & Transportation*, vol. 23, no. 1, pp. 11–17, 2011.

[31] S. Yu, Y. Jia, and D. Sun, "Identifying factors that influence the patterns of road crashes using association rules: a case study from Wisconsin, United States," *Sustainability*, vol. 11, no. 7, Article ID 1925, 2019.

[32] R. Rusli, M. M. Haque, M. Saifuzzaman, and M. King, "Crash severity along rural mountainous highways in Malaysia: an application of a combined decision tree and logistic regression model," *Traffic Injury Prevention*, vol. 19, no. 7, pp. 741–748, 2018.

[33] A. Pande and M. Abdel-Aty, "Discovering indirect associations in crash data through probe attributes," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2083, no. 1, pp. 170–179, 2008.

[34] J. Hong, J. Park, G. Lee, and D. Park, "Endogenous commercial driver's traffic violations and freight truck-involved crashes on mainlines of expressway," *Accident Analysis & Prevention*, vol. 131, pp. 327–335, 2019.

[35] J. Hong, R. Tamakloe, and D. Park, "A comprehensive analysis of multi-vehicle crashes on expressways: a double hurdle approach," *Sustainability*, vol. 11, no. 10, Article ID 2782, 2019.

[36] C. Xu, J. Bao, C. Wang, and P. Liu, "Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China," *Journal of Safety Research*, vol. 67, pp. 65–75, 2018.

[37] K. Lai and N. Cerpa, "Support vs. confidence in association rule algorithms," in *Proceedings of the OPTIMA Conference*, pp. 1–14, Curicó, Chile, October 2001.