Random Forest Screening and Association Rule Mining for Crash Pattern Discovery

Mohammad Reza Abbaszadeh Lima*
Independent Researcher
Ewing, NJ, 08628
Email: mr.abbaszadelima@gmail.com
ORCID ID: 0009-0008-7728-8811

Majid Rezaei
MBA Student, Allameh Tabataba'i University
Department of Management and Accounting
Allameh Tabataba'i University, Postal Code: 1489684511
Email: rezaei_majid@atu.ac.ir
ORCID ID: 0009-0002-4927-4419

Md Mahmud Hossain
Assistant Research Scientist
Texas A&M Transportation Institute, Bryan, TX, 77807
Email: M-Hossain@tti.tamu.edu
ORCID ID: 0000-0002-2737-6951

Ernest Nsong Asiedu
Graduate Research Assistant
Department of Civil and Environmental Engineering
Auburn University
Email: ena0011@auburn.edu
ORCID ID: 0009-0002-1029-9459

*Corresponding Author

ABSTRACT

We present a statistically principled pipeline for pattern discovery in crash data and apply it to fatal heavy vehicle crashes in rural Interstate work zones using NHTSA FARS (2020 to 2023). The two stage workflow first screens predictors of driver fatality at the scene via random forest variable importance, and second, mines interpretable association rules (Apriori) on the retained features. Rules are learned under minimum support = 0.03 and confidence = 0.60, and ranked by lift; the 16 highest lift rules are analyzed. The resulting patterns show consistent strength (lift = 3.18 to 3.85; mean confidence = 0.64; mean support = 3.46%) and most often combine disabling vehicle deformation, non collision first harmful events (run off road or fixed object), and roadside locations; lack of airbag deployment appears frequently among antecedents, while speed related items are rare. We comment on rule redundancy. The contribution is an end to end, reproducible applied statistics workflow that includes data preprocessing templates for FARS, integration of screening and rule mining, and interpretation guidelines, yielding transparent surrogates that complement black box severity models. Although the case study concerns work zone heavy vehicle crashes, the methodology is general and transferable to other domains requiring interpretable pattern mining. We also summarize the paper's main shortcomings to situate the results. These caveats are paired with suggested extensions so readers can assess scope and applicability.

Keywords: Association rule mining, Random forest, Crash data

INTRODUCTION

Work zones are important for maintaining and improving roads, but they can cause serious safety and traffic problems. This is especially true for heavy vehicles because they are large, heavy, and harder to drive (*1*). Their presence can also put other drivers at risk due to the changed traffic flow and tougher driving conditions.

Focusing on work zone crashes that lead to fatalities on Interstates is essential because this road type differs significantly in design, traffic patterns, and crash dynamics. Interstates are high-speed, limited-access highways with long-haul traffic and a lack of traffic signal devices, making them different in nature from other types of roads. Since other types of facilities have frequent intersections and more pedestrian or turning movements, a lower posted speed limit leads to different crash types. These differences affect driver behavior, risk exposure, and the nature of fatal crashes, specifically in work zones.

Examining rural Interstate work zones lets us evaluate how their higher speed limits, coupled with long, uninterrupted work-zone segments that keep drivers at highway speeds for miles, shape fatal-crash risk and guide targeted countermeasures. Since Interstate roads are built to higher standards than other road types and typically have no traffic signals, consistent pavement markings, proper lighting, and well-maintained surfaces, there are fewer hidden factors and confounding variables influencing crashes (*2*). This makes it easier to narrow down the factors involved in fatal crashes, compared to studies that include all types of roads. These crashes threaten drivers, workers, and other road users, leading to major safety issues and disruptions that hinder roadway efficiency.

The combination of work zone complexities and the unique operational demands of heavy vehicles, such as their size and weight, underscores the need for tailored safety measures to address crash characteristics specific to this vehicle type.

Understanding these key differences is crucial for developing targeted safety interventions. Moreover, research on this topic is limited, and addressing the gap can improve work zone safety strategies for this roadway vehicle type.

Research on work zone safety has thoroughly investigated crash characteristics, including severity, crash types, and patterns across time and space, to strengthen roadway safety strategies. Studies have shown that the majority of work zone crashes occur in the activity area, with rear-end collisions being the most common crash type across different highway classifications (*3*).

Heavy vehicle crashes in work zones are especially severe, with Interstates often seeing rollovers due to high speeds, while other types of work zones experience more rear-end collisions in congested settings. Crash rates during congested conditions in work zones are significantly higher, up to 24 times, than those observed under free-flow traffic, emphasizing the critical safety challenges posed by traffic queues in heavy vehicle operations (*4*).

Work zone setups, like narrowed lanes or temporary barriers, shape crash outcomes differently on Interstates and different types of roads, affecting aspects like vehicle damage or occupant injury patterns. Roadside barriers are widely recommended as a safety measure since collisions with barriers generally result in less severe injuries compared to rollovers or impacts with roadside hazards (*4*).

Association rule mining, a powerful data-driven approach, effectively uncovers patterns in crash characteristics, such as combinations of rollovers, airbag deployments, or vehicle damage, without relying on fixed assumptions (*5*). This method is ideal for identifying differences in crash outcomes, providing clear guidance for safety improvements (*5*). By focusing on crash characteristics rather than causes, it highlights how work zone designs create unique crash profiles, enabling agencies to prioritize measures.

The existence of work zones significantly elevates crash rates, with rear-end collisions being the most common crash type, and crash occurrences are unevenly distributed throughout different work zone areas (*6*).

Prior studies emphasize analyzing crash characteristics to inform effective safety policies, yet a detailed focus on heavy vehicle crashes in work zones across rural Interstates is scarce. The distinct challenges of heavy vehicles, including their size and maneuverability, amplify crash severity in work zones, particularly on high-speed Interstates compared to urban arterials.

Driver injury severity in large truck crashes is notably elevated in rural Interstate work zones, with lane-shift configurations and driver carelessness identified as critical risk factors (*7*). Empirical data indicate that heavy vehicles exhibit significantly lower free-flow speeds in work zones compared to passenger cars, despite identical posted speed limits (*8*). The economic and societal impacts of these crashes highlight the urgency of using advanced methods to understand and mitigate crash patterns.

This study addresses a crucial gap by utilizing association rule mining to examine the characteristics of heavy vehicle crashes in work zones on rural Interstates, drawing on FARS data from 2020–2023. The aim is to identify the key factors commonly present in these crash events.

These insights aim to guide transportation agencies in developing targeted safety measures to reduce fatalities and enhance safety for heavy vehicles. Additionally, the findings provide a foundation for policy decisions that improve work zone safety and operational efficiency across rural Interstate work zones. By leveraging advanced data analytics, this study seeks to advance the understanding of crash characteristics, offering a model for future research on heavy vehicle safety in work zones.

LITERATURE REVIEW

This section reviews prior studies on large-truck crashes, focusing on methods and key variables used to model crashes. While many researchers have studied truck-involved crashes, few have addressed work zone incidents involving heavy vehicles on rural Interstates. These roadways present unique crash conditions due to high speeds and limited access points. The combination of heavy trucks, work zones, and rural Interstates is rarely explored in the literature. This study aims to fill that gap using association rule mining on recent FARS data.

Study (*9*) investigated truck-involved crashes occurring in work zones across North Carolina using police-reported crash data from 2005 to 2014. Their analysis incorporated a range of variables, including driver, roadway, work-zone, environmental, and vehicle characteristics. To capture unobserved heterogeneity, they employed both mixed logit and partial proportional odds models. The findings revealed that the PPO model performed better and highlighted that the factors influencing injury severity varied between rural and urban highways. Injury severity was categorized into three levels: fatal/incapacitating/non-incapacitating, possible injury, and property damage only, and analyzed through separate models. Further statistical tests validated the need to estimate distinct models for rural and urban roadway settings.

Study (*10*) conducted in South Carolina analyzed truck-involved crashes in work zones by developing two separate mixed logit models based on roadway speed limits: one for non-Interstates with limits under 60 mph and another for Interstates with limits of 60 mph or higher. Using crash data from 2014 to 2020, the study identified several significant factors contributing to injury outcomes, including dark lighting conditions, female at-fault drivers, and driving too fast for roadway conditions, which were consistent across both models. Additionally, certain variables were found to be significant only under specific speed conditions; for example, sideswipe collisions, the presence of workers, and drivers under

35 years old were relevant for non-Interstate roads, while rear-end collisions, crashes involving three or more vehicles, and weekday occurrences were more significant on Interstates. Findings emphasized the need for separate models based on speed environments and suggested targeted interventions to improve safety, particularly in relation to speeding and low-visibility conditions. This work highlights the importance of context-specific modeling in understanding truck-related work zone crashes.

Study (*11*) conducted research in Florida on single-vehicle truck crashes from 2011 to 2019 and applied a random-parameters multinomial logit model to compare driver injury outcomes in work-zone versus non-work-zone segments. They found that crashes on rural Interstates carried up to a fourteen-fold higher risk of severe injury and that unique work-zone features, such as lane-shift layouts, and driver behaviors like speeding were key contributors to injury severity. These findings highlight how rural Interstate work zones differ fundamentally from other road contexts. They found that crashes on rural Interstates carried up to a fourteen-fold higher risk of severe injury and that unique work-zone features and driver behaviors like speeding were key contributors to injury severity.

Finding of (*12*) compared several econometric models to identify factors that drive injury severity in large-truck crashes within work zones in Minnesota. Their results showed that the generalized ordered logit framework fit the ordinal injury data best and that daytime crashes, higher speed limits, lack of access control, and rural principal arterial settings had the greatest impact on severe outcomes. By focusing on these causal factors in work-zone environments, this study highlights the importance of roadway context and traffic control features, insights we extend to rural Interstates through association rule mining of FARS data from 2020–2023.

Although these studies offer valuable insights into the factors influencing injury severity in work-zone truck crashes, none have explored the hidden combinations of fatal crash characteristics that co-occur on rural Interstates. To address this, our research applies association rule mining to FARS data from 2020–2023, uncovering the most frequent and influential patterns of heavy-vehicle work-zone crashes. By moving beyond single-factor models to a pattern-based analysis, we aim to reveal multi-variable interactions that can inform targeted safety interventions.

METHODS

Association Rule Mining

Association rule mining (ARM) is a data-mining method used to find frequent combinations of variable attributes that tend to occur together within an event (e.g., a fatal rural Interstate work zone crash) (*5*, *13*).

ARM reveals relationships among factors without assuming prior causal links, making it a useful tool for traffic safety analysis (*5*, *14*).

Recent studies have used ARM as a decision-support approach to derive rules from multidimensional crash databases for specific categories of variables (*5*, *15–17*).

Let $P = \{p1, p2, p3, \ldots, pn\}$ denote the crash database, and let each observation in P contain a subset of items (variable attributes) from the itemset $Q = \{q1, q2, q3, \ldots, qn\}$.
A rule is written $U \rightarrow V$, where $U, V \subseteq Q$ and $U \cap V = \emptyset$. U is the antecedent (left-hand side, LHS), and V is the consequent (right-hand side, RHS).

In an n-itemset rule, multiple items may appear in the antecedent. For example, a three-itemset rule is {weather condition = rain, surface condition = wet} → {work-zone presence = yes}. These mined rules express interdependencies among factors rather than direct causation.

Rules are evaluated by three measures: support (S), confidence (C), and lift (L) (*2*, *5*). Here, "support" measures how often the rule or pattern ($U \rightarrow V$) occurs in the full dataset, whereas "confidence" is the share of times that $U \rightarrow V$ recurs relative to the number of times U appears (*2*, *5*). The

third measure, “lift,” indicates how frequently the items co-occur as part of the same independent crash events (*2*, *5*). The parameter equations are given below, formulas 1 to 5 (*5*):

$$\sup(V) = \frac{(V')}{N} \tag{1}$$

$$\sup(U) = \frac{(U')}{N} \tag{2}$$

$$\sup(U \to V) = \frac{(U' \cap V')}{N} \tag{3}$$

$$\text{Confidence}(U \to V) = \frac{\sup(U \to V)}{\sup(U)} \tag{4}$$

$$\text{Lift}(U \to V) = \frac{\sup(U \to V)}{\sup(U) \times \sup(V)} \tag{5}$$

Here, N is the number of desired crashes, U′ equals the frequency of occurrences with U, V′ equals the frequency of occurrences with V, and (U′ ∩ V′) equals the frequency of occurrences with both U and V.

In association rule discovery, lift is crucial for gauging rule strength because it captures how much more often the antecedent(s) and consequent are observed together in the same crash context than would be expected under statistical independence (*5*, *18*). A lift value greater than 1 signals a positive association between U and V, whereas a value below 1 suggests a negative association (*19*, *20*). A lift value close to 1 indicates that U is independent of the likelihood of V (*2*, *13*).

Study Data

The Fatality Analysis Reporting System (FARS), maintained by the National Highway Traffic Safety Administration (NHTSA), is a nationally standardized database of fatal motor vehicle crashes in the United States, serving as a cornerstone for transportation safety research (*21*). FARS data are compiled annually from police reports, vehicle records, medical reports, and other sources, providing detailed information on crash circumstances, vehicles, and involved persons. The database’s structure allows researchers to track not only the occurrence of fatalities but also their timing, enabling time-sensitive analyses of crash outcomes(*22*). Additionally, FARS provides critical data on driver intoxication, including alcohol and drug involvement, making it a key resource for studying factors present in fatal crashes (*23*).

For this study, FARS data from 2020 to 2023 were used, integrating person, vehicle-level, accident-level, and distraction-level datasets to analyze driver outcomes in work zone crashes. The analysis focused on a curated set of variables capturing person-level, vehicle-level, and accident-level characteristics. Person-level variables included age group, sex, alcohol test results, airbag deployment, ejection status, restraint use, drug involvement, drinking status, and distraction. Vehicle-level variables encompassed body type, speed limit, number of lanes, ownership, hit-and-run status, road surface condition, speed-related factors, alignment, deformation, traffic control, profile, trafficway type, and model year. Accident-level variables included functional system, lighting conditions, weather, relation to roadway, rural/urban classification, work zone presence, time block, holiday status, intersection type, manner of collision, and weekday status. Crashes were filtered to include only those occurring in work zones on rural Interstate roads involving vehicles with a Gross Vehicle Weight Rating (GVWR) of 10,001–26,000 pounds.

Table 1 presents the descriptive statistics of the dataset, providing insights that help inform the selection of variables for the subsequent feature importance ranking analysis.

TABLE 1 Descriptive statistics of the study data

Variable	Category	Percent
BODY_TYPNAME	Cross Country/Intercity Bus	5
	Medium/heavy Pickup (GVWR greater than 10,000 lbs.)	4.6
	Medium/Heavy Vehicle Based Motor Home	0.8
	Other Bus Type	0.2
	School Bus	1.9
	Single-unit straight truck or Cab-Chassis (GVWR greater than 26,000 lbs.)	5
	Single-unit straight truck or Cab-Chassis (GVWR range 10,001 to 19,500 lbs.)	3.3
	Single-unit straight truck or Cab-Chassis (GVWR range 19,501 to 26,000 lbs.)	6.2
	Truck-tractor (Cab only, or with any number of trailing unit; any weight)	72
	Unknown medium/heavy truck type	1
ROLLOVERNAME	No Rollover	90
	Rollover	6.4
	Rollover, Tripped by Object/Vehicle	2.9
	Rollover, Unknown Type	0.6
VSPD_LIMNAME	40 MPH	0.6
	45 MPH	1.5
	50 MPH	0.6
	55 MPH	18.5
	60 MPH	13.7
	65 MPH	21.8
	70 MPH	27.6
	75 MPH	14.9
	80 MPH	0.6
	Reported as Unknown	0.2
VNUM_LANNAME	Five lanes	0.2
	Four lanes	2.7
	Not Reported	0.6
	One lane	7.1
	Three lanes	6
	Two lanes	83.4

OWNERNAME	Driver (in this crash) Not Registered Owner (Other Private Owner Listed)	7.5
	Driver (in this crash) was Registered Owner	7.1
	Not Applicable, Vehicle Not Registered	0.6
	Unknown	1
	Vehicle Registered as Commercial/Business/Company/Government Vehicle	80.3
	Vehicle Registered as Rental Vehicle	3.5
L_COMPLNAME	No valid license for this class vehicle	3.1
	Not licensed	0.6
	Unknown	6.6
	Unknown If CDL and/or CDL endorsement required for this vehicle	0.6
	Valid license for this class vehicle	89
HIT_RUNNAME	No	99.4
	Yes	0.6
VSURCONDNAME	Dry	92.3
	Snow	2.5
	Wet	5.2
AGE_GROUP	Middle Adult (25-45)	40.5
	Middle-Aged (46-65)	40.9
	Older (65+)	8.5
	Young (<25)	10.2
SPEEDRELNAME	No	83.2
	Reported as Unknown	2.7
	Yes, Exceeded Speed Limit	1.9
	Yes, Specifics Unknown	2.7
	Yes, Too Fast for Conditions	9.5
VALIGNNAME	Curve - Left	1
	Curve - Right	5.8
	Curve - Unknown Direction	0.4
	Curve Left	2.1
	Curve Right	3.3
	Not Reported	0.8
	Straight	86.5
SEXNAME	Female	8.1
	Male	90.7
	Not Reported	0.6
	Reported as Unknown	0.6
DEFORMEDNAME	Damage Reported, Extent Unknown	4.1
	Disabling Damage	61.4
	Functional Damage	11.6
	Minor Damage	10.6

	No Damage	2.3
	Not Reported	9.8
	Reported as Unknown	0.2
VTRAFCONNAME	No Controls	61.4
	Not Reported	9.1
	Other	6
	Other Regulatory Sign	2.7
	Reported as Unknown	0.4
	Warning Sign	20.3
VPROFILENAME	Downhill	7.5
	Grade, Unknown Slope	11.8
	Hillcrest	2.3
	Level	72.2
	Not Reported	2.9
	Sag (Bottom)	0.4
	Uphill	2.9
VTRAFWAYNAME	Entrance/Exit Ramp	0.4
	Not Reported	0.2
	Two-Way Divided, Unknown if Unprotected Median or Positive Median Barrier	1
	Two-Way, Divided, Positive Median Barrier	53.9
	Two-Way, Divided, Unprotected Median	40.7
	Two-Way, Not Divided	3.7
FUNC_SYSNAME	Interstate	100
LGT_CONDNAME	Dark - Lighted	1.9
	Dark - Not Lighted	20.7
	Dark - Unknown Lighting	1.2
	Dawn	0.8
	Daylight	71.2
	Dusk	3.9
	Reported as Unknown	0.2
DRUGSNAME	No (drugs not involved)	39.4
	Not Reported	53.5
	Reported as Unknown	4.1
	Yes (drugs involved)	2.9
WEATHERNAME	Clear	81.7
	Cloudy	7.9
	Fog, Smog, Smoke	2.3
	Not Reported	2.1

	Rain	3.3
	Severe Crosswinds	0.2
	Snow	2.5
REL_ROADNAME	Gore	0.4
	On Median	5.4
	On Roadside	6.6
	On Roadway	79.9
	On Shoulder	1.7
	Outside Trafficway	5.2
	Reported as Unknown	0.2
	Separator	0.6
RUR_URBNAME	Rural	100
MAN_COLLNAME	Angle	3.9
	Front-to-Front	2.9
	Front-to-Rear	61.2
	Other	0.6
	Rear-to-Side	1.2
	Sideswipe - Opposite Direction	0.6
	Sideswipe - Same Direction	4.6
	The First Harmful Event was Not a Collision with a Motor Vehicle in Transport	24.9
DOANAME	Died at Scene	18.9
	Died En Route	0.2
	Not Applicable	80.9
WRK_ZONENAME	Construction	66.6
	Maintenance	7.3
	Work Zone, Type Unknown	26.1
TIME_BLOCK	12 A.M.–6 A.M.	12
	12 P.M.–6 P.M.	42.7
	6 A.M.–12 P.M.	24.1
	6 P.M.–12 A.M.	21
	Unknown	0.2
IS_HOLIDAY	No	99.8
	Yes	0.2
DRINKINGNAME	No (Alcohol Not Involved)	52.9
	Not Reported	39.4
	Reported as Unknown	6.2
	Yes (Alcohol Involved)	1.5
DISTRACTNAME	Distracted by Outside Person, Object or Event	0.2

	Distraction (Distracted), Details Unknown	1
	Distraction/Inattention	1
	Eating or Drinking	0.2
	Inattention (Inattentive), Details Unknown	5.2
	Not Distracted	37.8
	Not Reported	43.6
	Other Distraction [Specify:]	0.2
	Other Mobile Phone Related	0.2
	Reported as Unknown if Distracted	7.7
	While Manipulating Mobile Phone	0.8
	While Talking or Listening to Mobile Phone	0.8
	While Using or Reaching for Device/Object Brought into Vehicle	0.4
	While Using Other Component/Controls Integral to Vehicle	0.8
DRIVER_ALONE	FALSE	31.1
	TRUE	68.9
AIR_BAGNAME	Deployed- Combination	1.5
	Deployed- Front	2.5
	Deployed- Side (door, seatback)	0.2
	Deployment- Unknown Location	1.9
	Not Deployed	88.4
	Not Reported	1.5
	Reported as Deployment Unknown	4.1
EJECTIONNAME	Not Ejected	93.2
	Not Reported	0.2
	Partially Ejected	1
	Reported as Unknown if Ejected	0.6
	Totally Ejected	5
REST_USENAME	Lap Belt Only Used	0.6
	None Used/Not Applicable	13.3
	Not Reported	1
	Other	0.2
	Reported as Unknown	12.9
	Restraint Used - Type Unknown	0.4
	Shoulder and Lap Belt Used	70.3
	Shoulder Belt Only Used	1.2
Weekday	No	16.8
	Yes	83.2

Figure 1 visualizes the distribution of rural Interstate work zone crashes across states from 2020 to 2023, helping identify geographic patterns in crash frequency. It supports variable selection by revealing whether states with more rural areas are more prone to such crashes, guiding further analysis.

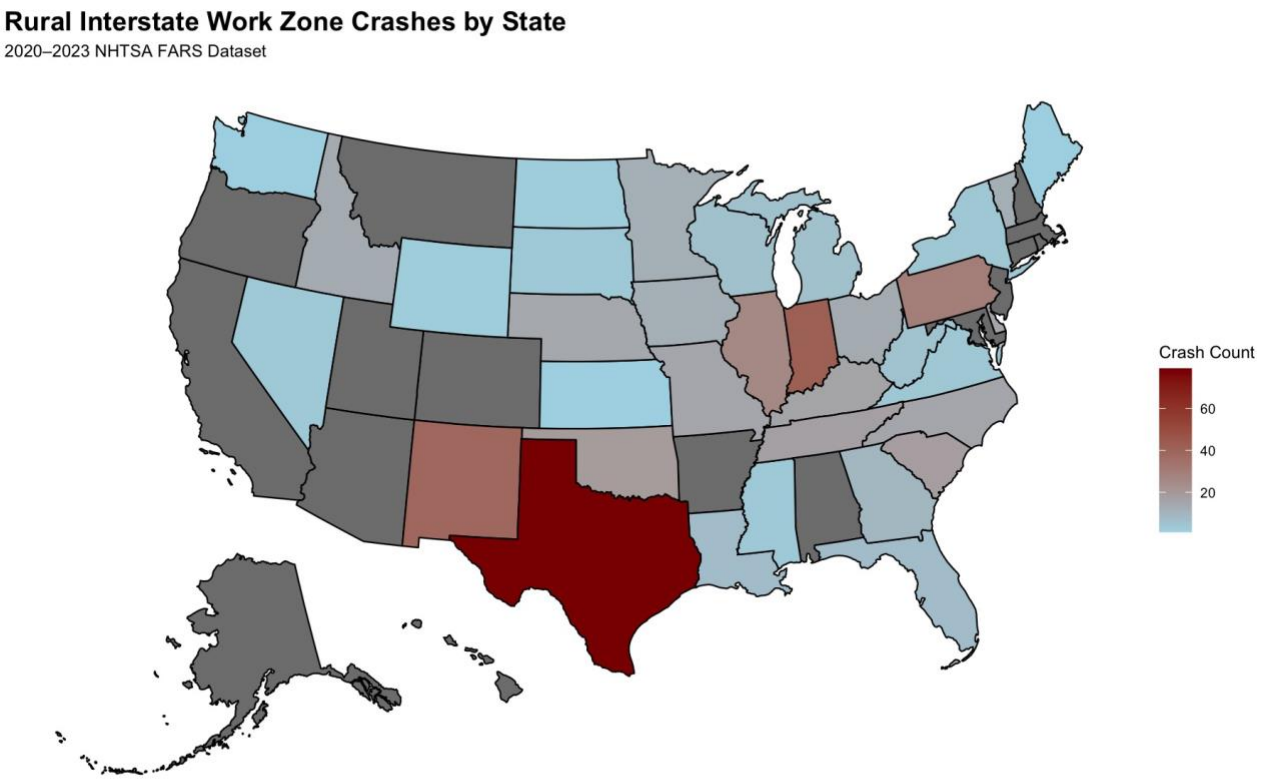

FIGURE 1 Choropleth map of fatal rural Interstate work zone crashes by state (NHTSA FARS 2020-2023)

Feature Selection

Data preprocessing, implemented via R scripts, involved merging datasets, handling missing values by coding them as "Unknown," and converting categorical variables to factors. Feature importance was assessed using a random forest model implemented with the ranger package in R, with driver outcome (DOANAME) as the response variable. The driver outcome variable (DOANAME) included categories such as Death on Arrival (DOA), which indicates fatalities pronounced dead at the crash scene.

The model, trained on 500 trees using the Gini impurity metric, ranked predictors based on their contribution to predicting driver outcomes.

Based on the feature importance ranking, the top variables contributing cumulatively to 80% of the total importance were selected for input into the association rule mining process. The graph below highlights these retained variables in orange, representing the first 17 variables in this case.

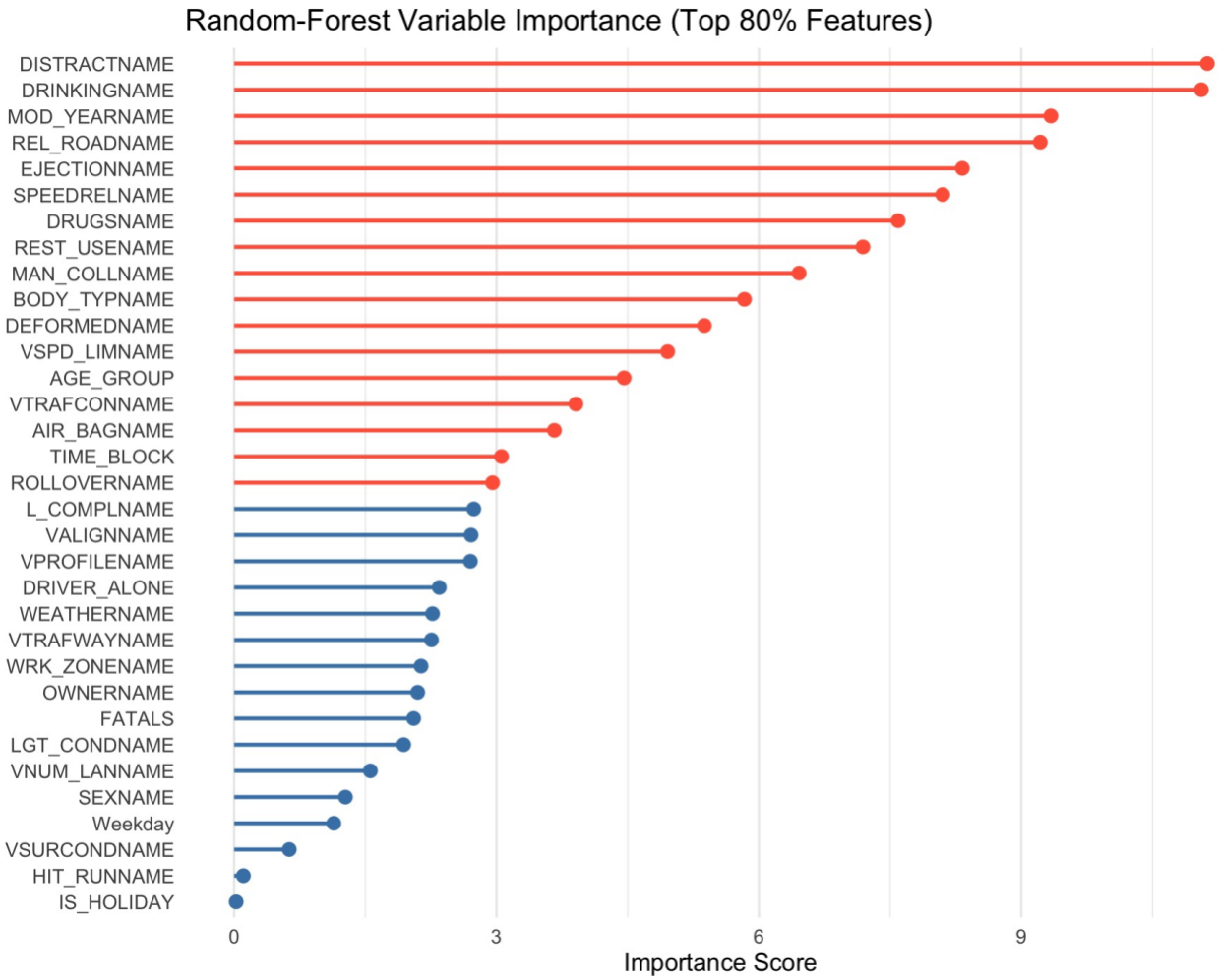

FIGURE 2 Random Forest variable importance (Top 80% features)

Association rule mining, conducted using the arules package, identified patterns linking predictors to DOANAME as the right-hand side for rural Interstate crashes, with a minimum support of 0.03 and confidence of 0.60.

RESULTS

This study utilized the Apriori algorithm on the 2020–2023 NHTSA FARS dataset of heavy vehicle work zone crashes on rural Interstates. The 16 rules with the highest lift values were retained and are presented in Table 2, sorted in descending order of lift to highlight the strongest associations between crash characteristics and fatal outcomes.

TABLE 2 Rural Interstate work zone rules

#	LHS	RHS	Support	Confidence	Lift
1	{EJECTIONNAME=Totally Ejected,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.033	0.727	3.852
2	{EJECTIONNAME=Totally Ejected}	{DOANAME=Died at Scene}	0.035	0.708	3.751

3	{REL_ROADNAME=On Roadside,DEFORMEDNAME=Disabling Damage,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.035	0.68	3.601
4	{REL_ROADNAME=On Roadside,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.035	0.68	3.601
5	{MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AGE_GROUP=Middle-Aged (46-65)}	{DOANAME=Died at Scene}	0.043	0.656	3.475
6	{REL_ROADNAME=On Roadside,DEFORMEDNAME=Disabling Damage}	{DOANAME=Died at Scene}	0.039	0.633	3.354
7	{REL_ROADNAME=On Roadside,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage}	{DOANAME=Died at Scene}	0.039	0.633	3.354
8	{REL_ROADNAME=On Roadside,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.035	0.629	3.334
9	{REL_ROADNAME=On Roadside,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.035	0.629	3.334
10	{MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,ROLLOVERNAME=Rollover}	{DOANAME=Died at Scene}	0.031	0.625	3.310
11	{DRUGSNAME=No (drugs not involved),MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.031	0.625	3.310
12	{MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AGE_GROUP=Middle-Aged (46-65),AIR_BAGNAME=Not Deployed}	{DOANAME=Died at Scene}	0.031	0.625	3.310
13	{REL_ROADNAME=On Roadside,EJECTIONNAME=Not Ejected,DEFORMEDNAME=Disabling Damage}	{DOANAME=Died at Scene}	0.033	0.615	3.259
14	{REL_ROADNAME=On Roadside,EJECTIONNAME=Not Ejected,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage}	{DOANAME=Died at Scene}	0.033	0.615	3.259
15	{EJECTIONNAME=Not Ejected,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AGE_GROUP=Middle-Aged (46-65)}	{DOANAME=Died at Scene}	0.035	0.607	3.215
16	{SPEEDRELNAME=No,MAN_COLLNAME=The First Harmful Event was Not a Collision with a Motor Vehicle in Transport,DEFORMEDNAME=Disabling Damage,AGE_GROUP=Middle-Aged (46-65)}	{DOANAME=Died at Scene}	0.031	0.6	3.178

Table 3 summarizes key metrics of the selected rules, showing that the rules have a high lift (3.18–3.85), high confidence (mean = 0.64), and moderate support (mean = 3.46%).

TABLE 3 A summary of the key metrics of association rules

Metric	Interstate (n = 16)
Lift range	3.18 – 3.85 (median = 3.33)
Mean support	3.46 %
Mean confidence	0.64

Table 4 lists the frequency of itemsets appearing on the left-hand side of the 16 association rules, with "Disabling Damage" and "Not a Collision with a Motor Vehicle in Transport" as the most common crash attributes.

TABLE 4 Interstate LHS Itemset frequency

Antecedent	Count
DEFORMEDNAME = Disabling Damage	11
MAN_COLLNAME = The First Harmful Event was Not a Collision with a Motor Vehicle in Transport	10
REL_ROADNAME = On Roadside	8
AIR_BAGNAME = Not Deployed	7
AGE_GROUP = Middle-Aged (46-65)	4
EJECTIONNAME = Not Ejected	3
EJECTIONNAME = Totally Ejected	2
ROLLOVERNAME = Rollover	1
DRUGSNAME = No (drugs not involved)	1
SPEEDRELNAME = No	1

The Apriori analysis on 2020–2023 rural Interstate work-zone crashes generated sixteen high-lift rules, all meeting a minimum support of 3% and confidence of 60%. On average, each rule covered 3.46% of the dataset and carried a confidence of 0.64, indicating that these antecedent combinations are both prevalent and reliably predictive of a fatal outcome. By focusing on the top-lift rules, we prioritized patterns that are most strongly associated with “DOANAME” while still maintaining reasonable coverage across the crash population.

Examining antecedent frequencies across those sixteen rules reveals that severe vehicle deformation (“DEFORMEDNAME = Disabling Damage”) appears in 11 rules, making it the single most common predictor of on-scene fatalities. Closely following are non-collision first

harmful events ("MAN_COLLNAME = The First Harmful Event was Not a Collision with a Motor Vehicle in Transport," 10 rules) and roadside crashes ("REL_ROADNAME = On Roadside," 8 rules). Other notable frequent antecedents include the absence of airbag deployment (7 rules) and the involvement of middle-aged drivers (4 rules). In contrast, rollover events, ejections, and drug-related impairment each feature only once or twice, suggesting that structural damage and run-off-road dynamics are the dominant fatal-crash association in these work zones in rural Interstates.

Co-occurrence analysis of the left-hand-side itemsets further underscores these relationships. Based on Table 5, the pair "Disabling Damage + Non-Collision First Harmful Event" occurs together in 8 of the 16 rules, and "Disabling Damage + On Roadside" appears in 6 rules, highlighting the critical intersection of high-severity damage and off-road departures. Tied at four occurrences each are combinations involving airbag nondeployment with either disabling damage, roadside location, or non-collision first events, while middle-aged drivers frequently co-occur with these structural and positional factors. These patterns point toward a fatal-crash profile dominated by off-road run-off scenarios with disabling damage.

TABLE 5 Top co-occurrences of itemsets

Rank	Items appearing together	Times present (out of 16)
1	DEFORMEDNAME = Disabling Damage + MAN_COLLNAME = The First Harmful Event was Not a Collision…	8
2	DEFORMEDNAME = Disabling Damage + REL_ROADNAME = On Roadside	6
3 (tie)	AIR_BAGNAME = Not Deployed + REL_ROADNAME = On Roadside	4
3 (tie)	AIR_BAGNAME = Not Deployed + DEFORMEDNAME = Disabling Damage	4
3 (tie)	AIR_BAGNAME = Not Deployed + MAN_COLLNAME = …Not a Collision…	4
3 (tie)	MAN_COLLNAME = …Not a Collision… + REL_ROADNAME = On Roadside	4
3 (tie)	AGE_GROUP = Middle-Aged (46–65) + DEFORMEDNAME = Disabling Damage	4
3 (tie)	AGE_GROUP = Middle-Aged (46–65) + MAN_COLLNAME = …Not a Collision…	4
3 (tie)	AGE_GROUP = Middle-Aged (46–65) + DEFORMEDNAME = Disabling Damage + MAN_COLLNAME = …Not a Collision…	4
4 (tie)	AIR_BAGNAME = Not Deployed + DEFORMEDNAME = Disabling Damage + MAN_COLLNAME = …Not a Collision…	3
4 (tie)	DEFORMEDNAME = Disabling Damage + MAN_COLLNAME = …Not a Collision… + REL_ROADNAME = On Roadside	3
5 (tie)	AIR_BAGNAME = Not Deployed + DEFORMEDNAME = Disabling Damage + REL_ROADNAME = On Roadside	2
5 (tie)	AIR_BAGNAME = Not Deployed + MAN_COLLNAME = …Not a Collision… + REL_ROADNAME = On Roadside	2
5 (tie)	DEFORMEDNAME = Disabling Damage + EJECTIONNAME = Not Ejected + REL_ROADNAME = On Roadside	2

Rank	Items appearing together	Times present (out of 16)
5 (tie)	DEFORMEDNAME = Disabling Damage + EJECTIONNAME = Not Ejected + MAN_COLLNAME = …Not a Collision…	2

DISCUSSIONS

The existing safety literature offers limited insight into work zone crashes involving heavy vehicles. This study addresses that gap by analyzing heavy vehicle crashes in work zones in rural Interstates using four years of data from the NHTSA FARS database.

To uncover hidden patterns and relationships, the study employs association rule mining, a data-driven method that does not rely on prior assumptions. This approach reveals distinct crash characteristics that account for the highest number of heavy vehicle crashes in the FARS dataset.

Among the 16 association rules with the highest lift, disabling vehicle damage, and the First Harmful Event was Not a Collision with a Motor Vehicle in Transport (e.g., rollovers or run-off-road crashes, crash with fixed object), stands out as a leading association of fatalities on Interstate roads.

These association rules covering 3.5 % of the sample on average, paint a remarkably consistent picture: fatal events are overwhelmingly characterized by severe vehicle deformation, single-vehicle off-road departures where the first harmful event is not a vehicle-to-vehicle collision, and the absence of occupant protection (air-bag non-deployment and ejection).

Alignment and contrasts with prior literature

Speed-related and multi-vehicle patterns

Study (*10*) reported that on South Carolina Interstates, injury risk rises with rear-end crashes, three-plus-vehicle involvement, and dark lighting. Our Interstate crash rules rarely include lighting or multi-vehicle factors, suggesting that once a truck leaves the travel lane, crash severity is driven mainly by impact force and roadside design rather than traffic conditions or visibility.

Rural/urban heterogeneity

Research study (*9*) showed that the lack of restraint uses and DUI dominate rural work-zone severity, whereas speeding is more noticeable in urban zones. Our findings echo the *restraint* result (air-bag non-deployment, ejection) but not the alcohol/speed factors, again pointing to crash-mechanism differences on rural Interstates where geometric departure, not driver excess speed, is crucial.

Single-vehicle crashes

Article (*24*) found driver-injury odds in single-truck crashes to be fourteen times higher on rural Interstates and six times higher on urban Interstate facilities than on comparable non-work-zone locations. Our association rules confirm the huge role of single-truck, infrastructure-interaction crashes, but add nuance by isolating disabling deformation and roadside location as the critical precursors.

Legacy evidence on work-zone configuration

Research study (*25*) identified higher posted speeds, two-way undivided alignments, and proximity to the activity area as the most important contributors to injury. While our dataset is

limited to divided Interstates, the prominence of on-roadside crashes in our rules suggests that the clear-zone area remains crucial even after two decades since this research was done.

Temporal instability and collision type
Research paper (*26*) used association rules to show that rear-end collisions dominate national fatal work-zone patterns and vary by time of day. In contrast, our rural-Interstate subset is dominated by non-collision harmful-events, implying that national rear-end trends may be driven by urban or mixed-traffic segments.

Speed-behavior implications
Both studies (*12*, *27*) emphasize high speed limits and traffic-control layout as key severity drivers. The near-absence of the speed-related item in our rule set (appearing only once) suggests that even normative speeds on 70-mph rural Interstates can produce fatal outcomes once a heavy vehicle departs the road.

CONCLUSIONS

This paper used association rule mining on four years of FARS data to study heavy-vehicle crashes in rural Interstate work zones. The rules with the highest lift show a clear pattern: many fatal events involve a first harmful event that is not a vehicle-to-vehicle crash, occur at the roadside, and result in disabling vehicle damage. Air bag non-deployment appears often among the rules, while speed-related items appear rarely.

These results argue for roadway-specific safety strategies. For rural Interstate work zones, the priority should be engineering measures that prevent run-off-road events and reduce impact severity when they occur. Examples include stronger channelization and taper design, temporary positive barriers or guardrails where feasible, wider buffer and clear zones, hazard shielding, and better roadside delineation. Since airbag non-deployment is common in the rules, programs that promote proper maintenance and deployment of restraint systems deserve attention. At the same time, departures from the travel lane can be fatal, so keeping heavy vehicles within the lane and away from fixed objects is key.

Prior studies point to restraint use and speed management on arterials; our Interstate findings complement that literature by pointing to run-off-road mechanisms rather than multi-vehicle conflicts. Together, these strands support tailored policies by facility type rather than a single policy for all work zones.

LIMITATIONS

This study is associative and does not claim causation. The analysis relies on FARS fatal-crash records (2020–2023), so results reflect conditions present in crashes that resulted in a death rather than risk across all crash types. The period also includes pandemic-era travel and enforcement shifts that may introduce temporal confounding.

The scope is intentionally narrow, so generalization to urban facilities or arterial contexts should be made with caution. Several variables are subject to measurement error or reporting bias, and missingness handled via “Unknown” categories can create artifacts that appear in rules. Future work should add arterial results, link to the detailed work zone layout. Analytical choices (e.g., coding strategies, thresholds, and model settings) follow established practice, though alternative reasonable specifications could yield similar, directionally consistent insights. The

findings are best interpreted as decision-support signals that complement other methodologies and datasets.

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows:
Study conception and design: Mohammad Reza Abbaszadeh Lima, Majid Rezaei
Data collection: Mohammad Reza Abbaszadeh Lima, Majid Rezaei, Ernest Nsong Asiedu
Analysis and interpretation of results: Mohammad Reza Abbaszadeh Lima, Majid Rezaei, Md Mahmud Hossain
Draft manuscript preparation: Mohammad Reza Abbaszadeh Lima, Majid Rezaei, Md Mahmud Hossain, Ernest Nsong Asiedu
All authors reviewed the results and approved the final version of the manuscript.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research and authorship of this article.

FUNDING

The authors disclosed no financial support for the research and authorship of this article.

REFERENCES

1. Haque, M. S., L. R. Rilett, and L. Zhao. Impact of Platooning Connected and Automated Heavy Vehicles on Interstate Freeway Work Zone Operations. *Journal of Transportation Engineering, Part A: Systems*, Vol. 149, No. 3, 2023. https://doi.org/10.1061/JTEPBS.TEENG-7434.

2. Hossain, M. M., M. R. A. Lima, and H. Zhou. Potential Impacts of Connected and Autonomous Vehicles on Controlling Criteria for Road Geometric Design: A Review. *Transportation Research Record: Journal of the Transportation Research Board*, 2024. https://doi.org/10.1177/03611981241242764.

3. Garber, N. J., and M. Zhao. Distribution and Characteristics of Crashes at Different Work Zone Locations in Virginia. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1794, No. 1, 2002, pp. 19–25. https://doi.org/10.3141/1794-03.

4. Mekker, M. M., S. M. Remias, M. L. McNamara, and D. M. Bullock. *Characterizing Interstate Crash Rates Based on Traffic Congestion Using Probe Vehicle Data*. West Lafayette, IN, 2020.

5. Hossain, M. M., M. Lima, and H. Zhou. Severity Analysis of Secondary Crashes on High-Speed Roadways: Pattern Recognition Using Association Rule Mining. *Transportation Research Record Journal of the Transportation Research Board*, 2024. https://doi.org/https://doi.org/10.1177/03611981231223194.

6. Yang, H., K. Ozbay, O. Ozturk, and K. Xie. Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. *Traffic Injury Prevention*, Vol. 16, No. 4, 2015, pp. 387–396. https://doi.org/10.1080/15389588.2014.948615.

7. Islam, M. An Empirical Analysis of Driver Injury Severities in Work-Zone and Non-Work-Zone Crashes Involving Single-Vehicle Large Trucks. *Traffic Injury Prevention*, Vol. 23, No. 7, 2022, pp. 398–403. https://doi.org/10.1080/15389588.2022.2101643.

8. Chitturi, M. V., and R. F. Benekohal. Effect of Lane Width on Speeds of Cars and Heavy Vehicles in Work Zones. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1920, No. 1, 2005, pp. 41–48. https://doi.org/10.1177/0361198105192000105.

9. Yu, M., C. Ma, C. Zheng, Z. Chen, and T. Yang. Injury Severity of Truck-Involved Crashes in Work Zones on Rural and Urban Highways: Accounting for Unobserved Heterogeneity. *Journal of Transportation Safety & Security*, Vol. 14, No. 1, 2022, pp. 83–110. https://doi.org/10.1080/19439962.2020.1726544.

10. Madarshahian, M., A. Balaram, F. Ahmed, N. Huynh, C. K. A. Siddiqui, and M. Ferguson. Analysis of Injury Severity of Work Zone Truck-Involved Crashes in South Carolina for Interstates and Non-Interstates. *Sustainability*, Vol. 15, No. 9, 2023, p. 7188. https://doi.org/10.3390/su15097188.

11. Al-Bdairi, N. S. S., and S. Hernandez. An Empirical Analysis of Run-off-Road Injury Severity Crashes Involving Large Trucks. *Accident Analysis & Prevention*, Vol. 102, 2017, pp. 93–100. https://doi.org/10.1016/j.aap.2017.02.024.

12. Osman, M., R. Paleti, S. Mishra, and M. M. Golias. Analysis of Injury Severity of Large Truck Crashes in Work Zones. *Accident Analysis & Prevention*, Vol. 97, 2016, pp. 261–273. https://doi.org/10.1016/j.aap.2016.10.020.

13. Hossain, M. M., X. Sun, E. Mitran, and M. A. Rahman. Investigating Fatal and Injury Crash Patterns of Teen Drivers with Unsupervised Learning Algorithms. *IATSS Research*, Vol. 45, No. 4, 2021. https://doi.org/10.1016/j.iatssr.2021.07.002.

14. Pande, A., 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, 2009. https://doi.org/10.1016/j.ssci.2007.12.001.

15. Abbaszadeh Lima, M. R., M. M. Hossain, H. Zhou, and Y. Song. Data Mining Approach to Explore the Contributing Factors to Fatal Wrong-Way Crashes by Local and Non-Local Drivers. *Future Transportation*, Vol. 4, No. 3, 2024, pp. 985–999. https://doi.org/10.3390/futuretransp4030047.

16. Hossain, M. M., H. Zhou, M. A. Rahman, S. Das, and X. Sun. Cellphone-Distracted Crashes of Novice Teen Drivers: Understanding Associations of Contributing Factors for Crash Severity Levels and Cellphone Usage Types. *Traffic Injury Prevention*, Vol. 23, No. 7, 2022. https://doi.org/10.1080/15389588.2022.2097667.

17. Hossain, M. M., H. Zhou, and S. Das. Data Mining Approach to Explore Emergency Vehicle Crash Patterns: A Comparative Study of Crash Severity in Emergency and Non-Emergency Response Modes. *Accident Analysis & Prevention*, Vol. 191, 2023, p. 107217. https://doi.org/10.1016/j.aap.2023.107217.

18. Kong, X., S. Das, K. Jha, and Y. Zhang. Understanding Speeding Behavior from Naturalistic Driving Data: Applying Classification Based Association Rule Mining. *Accident Analysis & Prevention*, Vol. 144, 2020, p. 105620. https://doi.org/10.1016/j.aap.2020.105620.

19. Das, S., R. Tamakloe, H. Zubaidi, I. Obaid, and A. Alnedawi. Fatal Pedestrian Crashes at Intersections: Trend Mining Using Association Rules. *Accident Analysis & Prevention*, Vol. 160, 2021, p. 106306. https://doi.org/10.1016/j.aap.2021.106306.

20. Samerei, S. A., K. Aghabayk, N. Shiwakoti, and S. Karimi. Modelling Bus-Pedestrian Crash Severity in the State of Victoria, Australia. *International Journal of Injury Control and Safety Promotion*, Vol. 28, No. 2, 2021, pp. 233–242. https://doi.org/10.1080/17457300.2021.1907597.

21. Tormoehlen, S., and J. M. Rudolphi. Summary of Roadway Incidents Involving Farm Equipment in Five Midwestern States Using the Fatality Analysis Reporting System (FARS). *Journal of Agromedicine*, Vol. 29, No. 3, 2024, pp. 504–507. https://doi.org/10.1080/1059924X.2024.2333552.

22. Yasmin, S., N. Eluru, and A. R. Pinjari. Pooling Data from Fatality Analysis Reporting System (FARS) and Generalized Estimates System (GES) to Explore the Continuum of Injury Severity Spectrum. *Accident Analysis & Prevention*, Vol. 84, 2015, pp. 112–127. https://doi.org/10.1016/j.aap.2015.08.009.

23. Walters, J. K., K. K. Repp, and M. C. Mew. Alcohol and Drug Presence in Traffic Crash Fatalities before and after the COVID-19 Pandemic: Evaluation of the Fatality Analysis Reporting System (FARS) and Linked Medical Examiner-Vital Records Data in Clackamas, Multnomah, and Washington County, Oregon, 2019–2021. *Forensic Science International: Synergy*, Vol. 8, 2024, p. 100468. https://doi.org/10.1016/j.fsisyn.2024.100468.

24. Islam, M. An Empirical Analysis of Driver Injury Severities in Work-Zone and Non-Work-Zone Crashes Involving Single-Vehicle Large Trucks. *Traffic Injury Prevention*, Vol. 23, No. 7, 2022, pp. 398–403. https://doi.org/10.1080/15389588.2022.2101643.

25. Bligh, R. P. Determining Design Wind Loads for Work Zone Traffic-Control Devices. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1877, No. 1, 2004, pp. 117–125. https://doi.org/10.3141/1877-13.

26. Das, S., A. Dutta, R. Tamakloe, and M. N. Khan. Analyzing the Time-Varying Patterns of Contributing Factors in Work Zone-Related Crashes. *Journal of Transportation Safety & Security*, Vol. 16, No. 6, 2024, pp. 655–682. https://doi.org/10.1080/19439962.2023.2246020.

27. Porter, R. J., and J. M. Mason. Modeling Speed Behavior of Passenger Cars and Trucks in Freeway Construction Work Zones: Implications on Work Zone Design and Traffic Control Decision Processes. *Journal of Transportation Engineering*, Vol. 134, No. 11, 2008, pp. 450–458. https://doi.org/10.1061/(ASCE)0733-947X(2008)134:11(450).