Applying Advanced Techniques to Datamine Pedestrian Crash Data

NOVEMBER 2023

Prepared By:

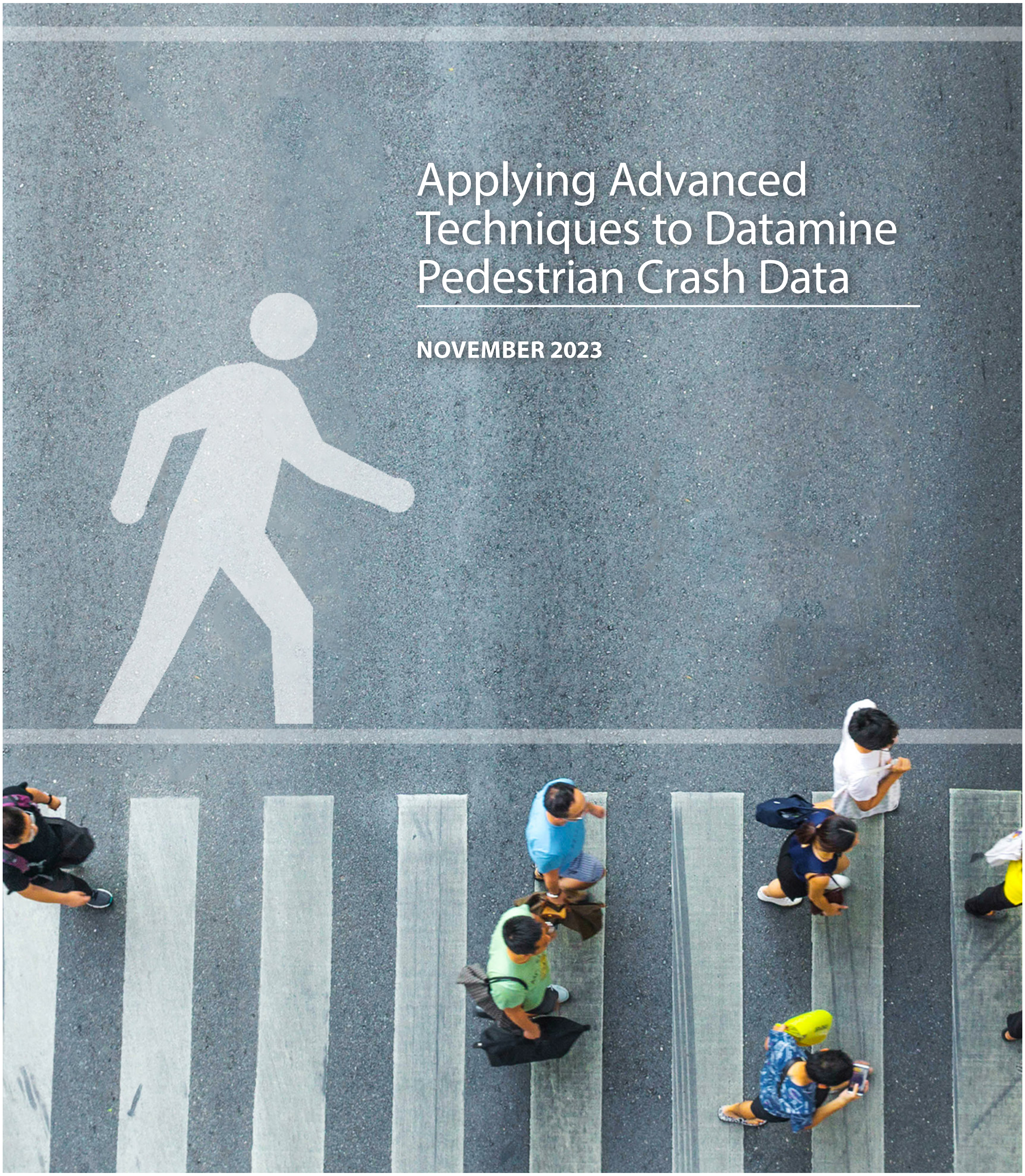
Texas A&M Transportation Institute

Prepared For:
Texas A&M Transportation Institute's
Center for Transportation Safety

TTI | Center for Transportation Safety

Minh Le, Research Engineer
Michael P. Pratt, Associate Research Engineer
Amir Hossein Oliaee, Ph.D. Student
Subasish Das, Associate Research Scientist
Mahin Ramezani, Research Data Scientist
Jason Wu, Associate Research Scientist
Sky Guo, Ph.D. Student

NOVEMBER 2023



Applying Advanced Techniques to Datamine Pedestrian Crash Data

NOVEMBER 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Pedestrian fatalities are increasing. We need to find out why and identify proper treatments.

INTRODUCTION

According to the National Highway Traffic Safety Administration (NHTSA), more than 6600 pedestrians died in traffic crashes in 2020 in the United States. A pedestrian was killed every 85 minutes, equivalent to around 119 people a week on average (1). Pedestrians are 1.5 times more likely to be killed in a crash on each trip than passenger vehicle occupants (2). Figure 1 presents a distribution of pedestrian fatalities as a percentage of total motor vehicle fatalities from 2008 to 2020. Table 1 shows that Texas has experienced a higher increase in the percentage of pedestrian fatalities than in the US since 2015. In fact, Figure 2 shows that pedestrian fatal crashes, as compared to total fatal crashes, has increased from 16% (2012) to 20% (2021) in Texas.

According to the National Highway Traffic Safety Administration (NHTSA), more than 6600 pedestrians died in traffic crashes in 2020 in the United States.

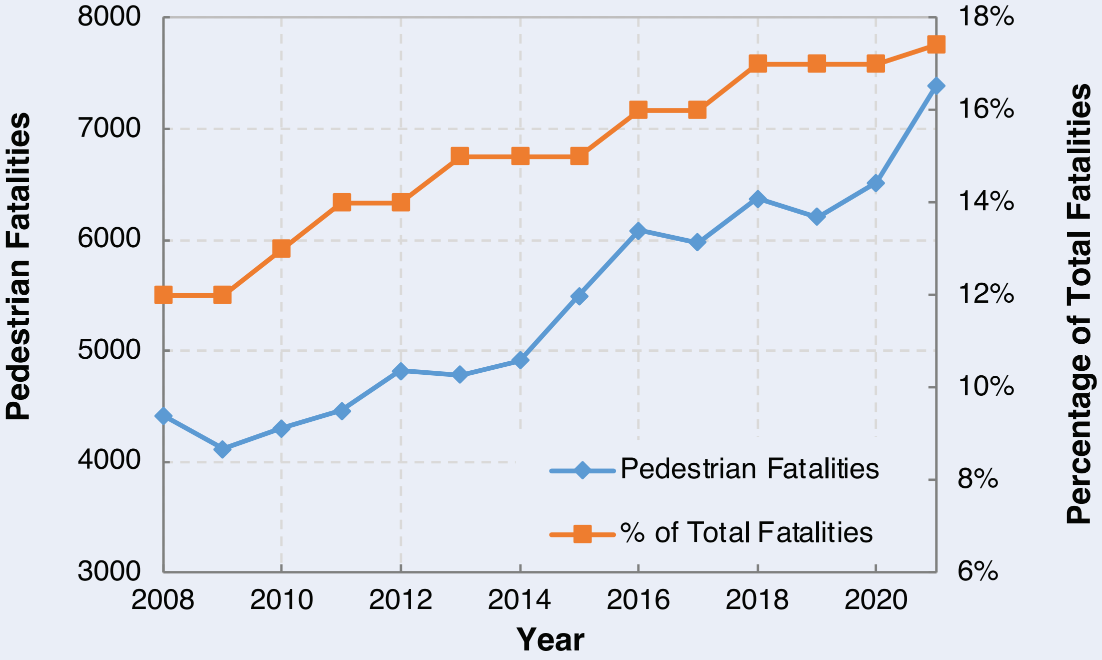


Figure 1. Pedestrian Fatalities in United States, 2008-2020.

There has been more focus on the safety of vulnerable roadway users, particularly pedestrians, however, most crash databases do not contain detailed information about pedestrians involved in crashes.

Table 1. US vs. Texas Pedestrian Fatalities (3).

YEAR	NUMBER OF FATALITIES (US)	NUMBER OF FATALITIES (TEXAS)
2015	5,494	549
2016	6,080	675
2017	6,075	608
2018	6,374	616
2019	6,272	649
2020	6,516	687
Difference 2020-2015	1,022	138
Percent Change	19%	25%

With these increasing trends, there has been more focus on the safety of vulnerable roadway users, particularly pedestrians, recently. However, most crash databases do not contain detailed information about pedestrians involved in crashes. Specifically, they don't provide the pedestrian actions or maneuvers at the time of the crash (pre-crash actions) or the intentionality of the person involved (i.e., was the person there intending to be a pedestrian or not). This level of information is needed to determine the appropriate countermeasures (if any) and yet is only often found in the narratives and diagrams of a crash report.

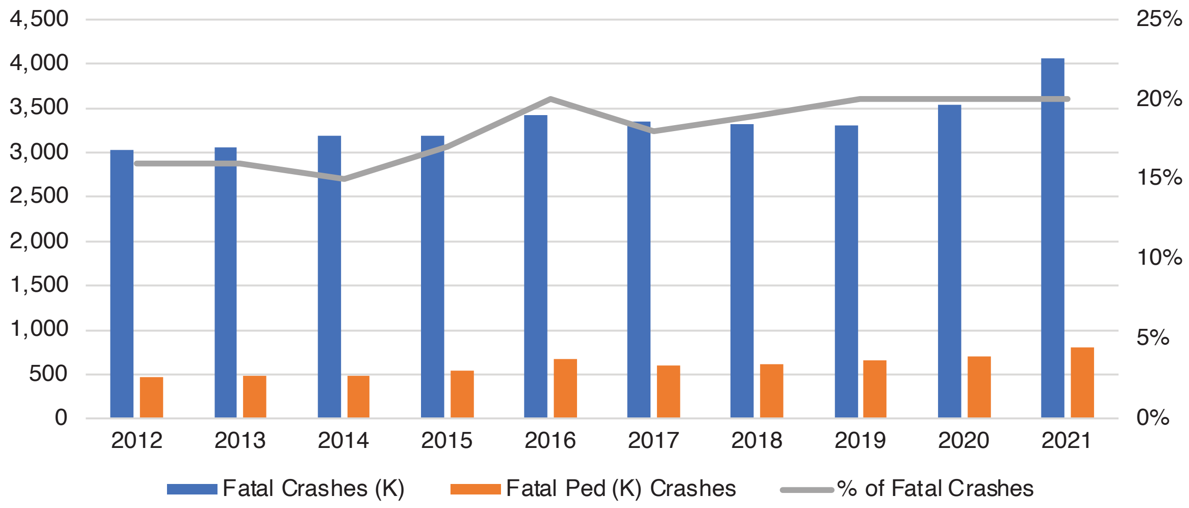


Figure 2. Texas Fatal Crashes vs. Fatal Pedestrian Crashes.

Analysists typically have to review the crash narratives and/or diagrams to determine the pre-crash actions or the ‘crash type’, as defined in FHWA’s Pedestrian and Bicycle Crash Analysis Tool (PBCAT) (4). They also need to infer the pedestrian’s intentionality by reviewing the crash circumstances (i.e., why was the pedestrian there?). These investigations can require substantial time and effort when analyzing a large dataset such as when analyzing a large area or a long time period. This information is not traditionally part of the ‘structured’ crash data which tends to be more centered around the motorized vehicles involved.

The purpose of this project is to explore advanced techniques to mine crash data for this type of information more efficiently. Ultimately, it can be used to develop a more robust crash database for alternative travel modes, starting with pedestrians. This crash database will supplement the ‘structured’ data with the ‘unstructured’ data, or interpreted fields, such as crash types and intentionality. Because weather often plays a role in mode choice, it will also include weather data, including sun glare determination, for each crash location and time. The same process can be expanded for other modes such as bicycles, e-scooters, wheelchairs, etc. The goal is to establish a robust alternative crash type repository for researchers and practitioners to use, explore, and advance the safety of vulnerable roadway users.

LITERATURE REVIEW

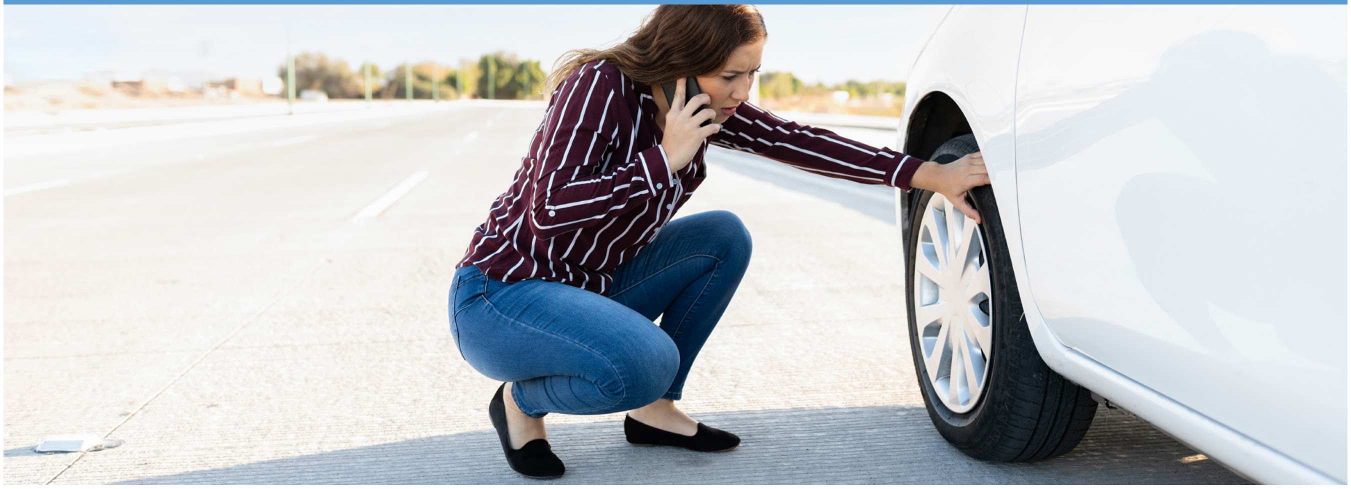
Pedestrian safety has been extensively studied, leading to a deeper comprehension of the factors influencing crashes. This substantial body of research forms the basis for addressing challenges of pedestrian safety like the ambiguity between pedestrian and vehicle interactions, intentionality, and related weather information.

This substantial body of research forms the basis for addressing challenges of pedestrian safety like the ambiguity between pedestrian and vehicle interactions, intentionality, and related weather information.

Crash Data Analysis Details

In 2000, the Federal Highway Administration (FHWA) and the National Highway Traffic Safety Administration (NHTSA) launched the PBCAT software, designed to aid in the analysis of non-motorist crashes and facilitate the identification of solutions and countermeasures(4). PBCAT was developed to describe the actions leading to traffic crashes between vehicles and pedestrians or bicyclists, allowing for a better understanding of event sequences and contributing efforts. The most recent version, PBCAT Version 3.0, has been updated to provide traffic safety professionals with enhanced knowledge of causation patterns, enabling the design of more effective interventions. This tool offers pre-drawn diagrams and dropdown menus, facilitating the collection of critical information from police reports about pedestrian/bicycle crashes (5). Despite the progress made with PBCAT, crash narrative insufficiency remains a concern, prompting exploration into the effectiveness of current narrations and the use of language models to improve understanding and comprehension. The relevance of PBCAT-related analysis has grown in the transportation safety community, with various studies using this methodology to examine pedestrian and bicyclist crash data. Shah et al. (6) investigated e-scooter and bicycle crashes, uncovering significant differences in multiple factors. Lopez et al. (7) evaluated the data quality of text narratives in police reports on bicycle crashes, revealing a high degree of missing information. Schneider and Stefanich (8) introduced the location-movement classification method (LMCM) and demonstrated its usefulness in providing insights not captured by PBCAT. Chavis et al. (9) analyzed pedestrian and bicycle crashes in Washington, D.C. and identified NHTSA crash groups, examining relevant countermeasures.

> The goal is to establish a robust alternative crash type repository for researchers and practitioners to use, explore, and advance the safety of vulnerable roadway users.



Overall, large language models and text mining analytics play a vital role in improving pedestrian safety research, enabling a more comprehensive understanding of crash factors and aiding in the development of effective interventions to reduce pedestrian-related crashes.

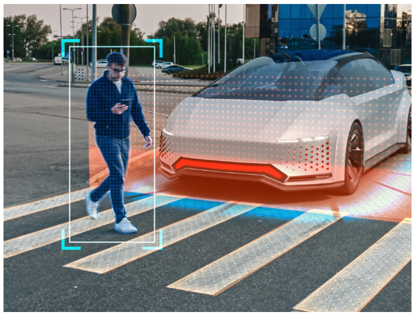
Due to the limited detail in the state or national crash databases, crash narratives play a crucial role in providing deeper insights into factors contributing to crash occurrences (10). However, manually identifying PBCAT-coded crash types from these extensive textual datasets is labor-intensive. Thus, text mining analytics offer a solution to extract insights from crash-narrative textual data efficiently (11). Various applications, such as thematic analysis, content analysis, supervised and unsupervised modeling, and natural language processing, have been utilized. Machine-learning models like XGBoost and Bidirectional Encoder Representations from Transformers (BERT) have successfully classified crash types and severity, demonstrating their promise in handling large language models for more accurate and efficient analysis (12-16). Researchers have also combined crash narratives with metadata to discern prevalent themes contributing to crash incidents (17). Overall, large language models and text mining analytics play a vital role in improving pedestrian safety research, enabling a more comprehensive understanding of crash factors and aiding in the development of effective interventions to reduce pedestrian-related crashes.

Computer Vision

Computer Vision (CV) has emerged as a powerful tool in the field of transportation safety, enabling researchers to extract valuable insights from complex datasets, such as pedestrian crash data. This section reviews the evolution of CV algorithms, with a focus on models like YOLO (You Only Look Once) and their applications in crash and safety analysis.

The field of CV has witnessed significant advancements over the past few decades. Early CV techniques relied on traditional methods like edge detection and feature extraction. However, recent years have seen a paradigm shift towards deep learning-based approaches. Convolutional Neural Networks (CNNs), in particular, have revolutionized the way we process and analyze visual data. These deep learning models are capable of automatically learning intricate patterns and features from images, making them well-suited for complex tasks such as object detection and recognition.

Convolutional Neural Networks (CNNs), in particular, have revolutionized the way we process and analyze visual data.



YOLO and similar models play a pivotal role in identifying and tracking pedestrians and other relevant objects in video footage or image data.

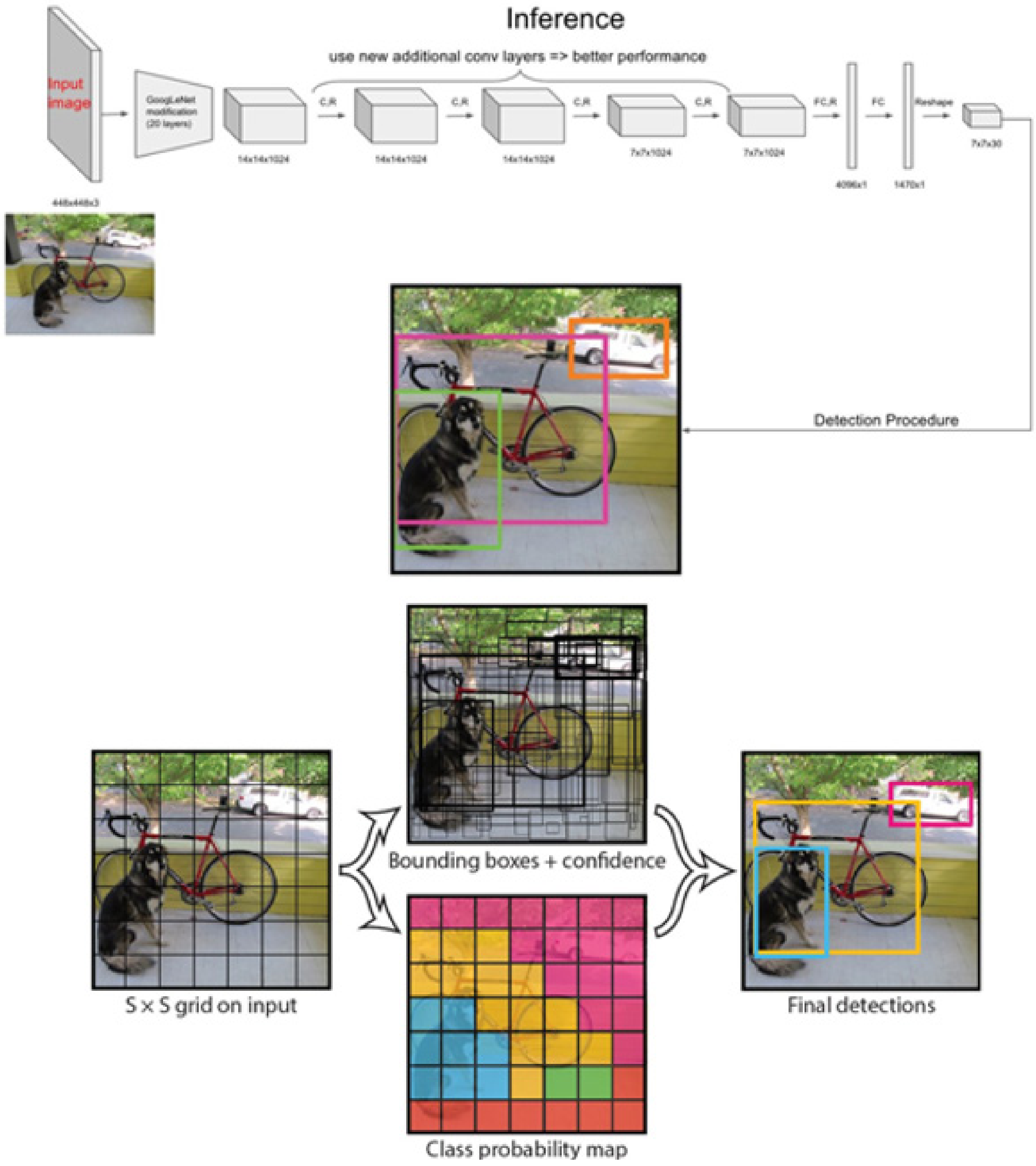
Among the many deep learning-based CV models, YOLO has gained considerable attention for its real-time object detection capabilities. YOLO's acronym, "You Only Look Once," encapsulates its efficiency in simultaneously predicting multiple objects in an image with remarkable speed. YOLO variants, such as YOLOv3 and YOLOv4, have further improved object detection accuracy and performance. YOLOv5 (18), focused on optimizing YOLO's architecture for real-time object detection while reducing computational complexity. It introduced a streamlined architecture with a focus on speed and efficiency and quickly gained popularity with its balance between speed and accuracy. The latest official YOLO version is YOLOv7 (19), which was released in January 2023 by the original authors of the YOLO architecture. According to them, YOLOv7 is the fastest and most accurate object detection algorithm available today. Figure 3 illustrates the YOLO network architecture (20) on an example picture that contains three objects to be detected – a dog, a bicycle, and a vehicle.

In the context of pedestrian crash data analysis, YOLO and similar models play a pivotal role in identifying and tracking pedestrians and other relevant objects in video footage or image data. This ability enables researchers to gather detailed information about pedestrian behavior, vehicle interactions, and environmental factors leading up to crashes.

Besides YOLO, there are many other CV models that have been developed and improved over the years. Some examples are:

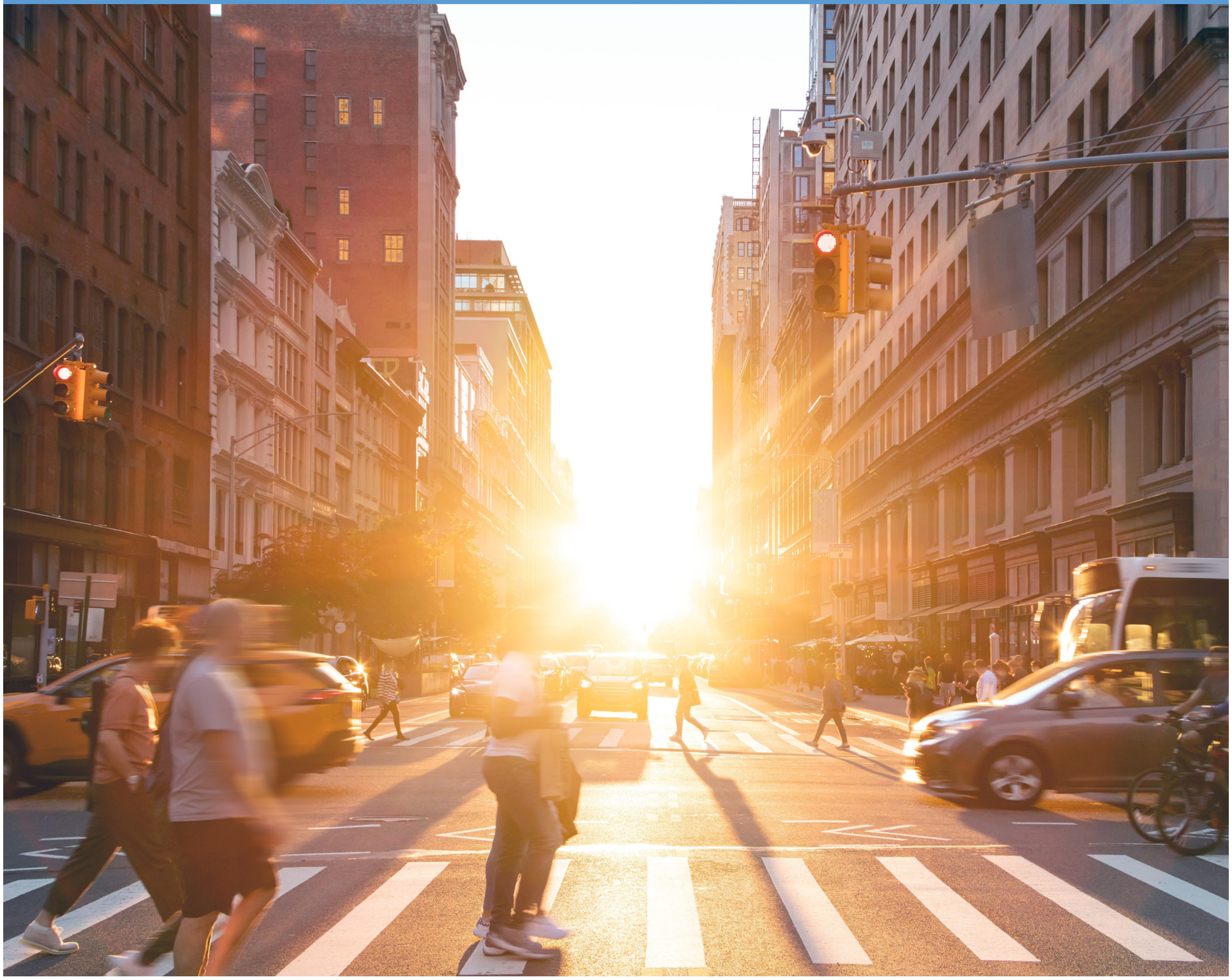
- **ResNet:** A deep convolutional neural network that uses residual connections to overcome the problem of vanishing gradients and achieve high accuracy on image classification tasks (21).
- **Mask R-CNN:** An extension of Faster R-CNN that adds a branch for predicting segmentation masks for each detected object, enabling instance segmentation and pixel-level object localization (22).
- **GAN:** A generative adversarial network that consists of two competing models: a generator that tries to produce realistic images from random noise, and a discriminator that tries to distinguish between real and fake images. GANs can be used for image synthesis, style transfer, super-resolution, and more (23).

Figure 3. YOLO Network Architecture (20).



The applications of CV in crash and safety analysis are multifaceted. CV algorithms can be utilized for:

- **Pedestrian Detection:** CV models excel at identifying pedestrians in various traffic scenarios, helping researchers understand pedestrian movements and the factors contributing to pedestrian-involved crashes. For example, Narayanan et al. (24) proposed a model that can accomplish pedestrian detection automatically using Histogram of Gradient (HOG) and YOLO. The paper also designed an alarm to alert the user on sight of pedestrians at night. Another study (25) used video data and physics-based simulation to reconstruct pedestrian crashes. The paper extracted the motion parameters of the vehicle and the pedestrian from the video data using CV techniques.
- **Traffic Flow Analysis:** CV-based traffic analysis can provide insights into traffic patterns, congestion, and potential bottlenecks, helping authorities optimize road infrastructure for pedestrian safety. Deng et al. (26) introduced a new dataset for traffic flow segmentation, which contains 1000 videos captured by 10 cameras in different locations and scenarios. The study also provided pixel-level annotations for vehicles and pedestrians, as well as vehicle counts and speeds. The study demonstrated that the CV model and the dataset can be used for various tasks, such as vehicle detection, tracking, counting, and speed estimation.
- **Anomaly Detection:** CV can be employed to detect unusual behaviors or events on roadways, such as jaywalking or sudden stops, which may be indicative of safety hazards. For example, Doshi and Yilmaz (27) proposed an efficient approach for a video anomaly detection system which is capable of running on roadway edge devices, e.g., on a roadside camera. Yuan et al. (28) proposed a novel method for anomaly detection in traffic scenes by reconstructing the motion of each pixel in a video frame using a spatial-aware convolutional neural network. The paper showed that the reconstruction error can be used as an anomaly score to detect abnormal events, such as collisions, illegal turns, and pedestrian crossings.



Weather

One of the challenges with studying the role of sun glare in crash trends is to identify crashes that may be induced by sun glare. Generally, bright sunlight can temporarily blind drivers when the sun sits at a relatively low altitude (usually just after sunrise or before sunset when the sun is near the horizon) and when the sun's rays parallel a driver's line of sight (29). In Mitra (30), crashes were defined as being caused by sun glare based on the vehicles' travel directions and time of day, defining morning and evening glare time windows for each month using National Oceanographic and Atmospheric Administration (NOAA) data. Work such as Ma et al. (29) and Hagita and Mori (31) applied a similar methodology. Due to the commonality between these studies, the researchers elected to use the benchmarks established in these prior analyses to guide their research.

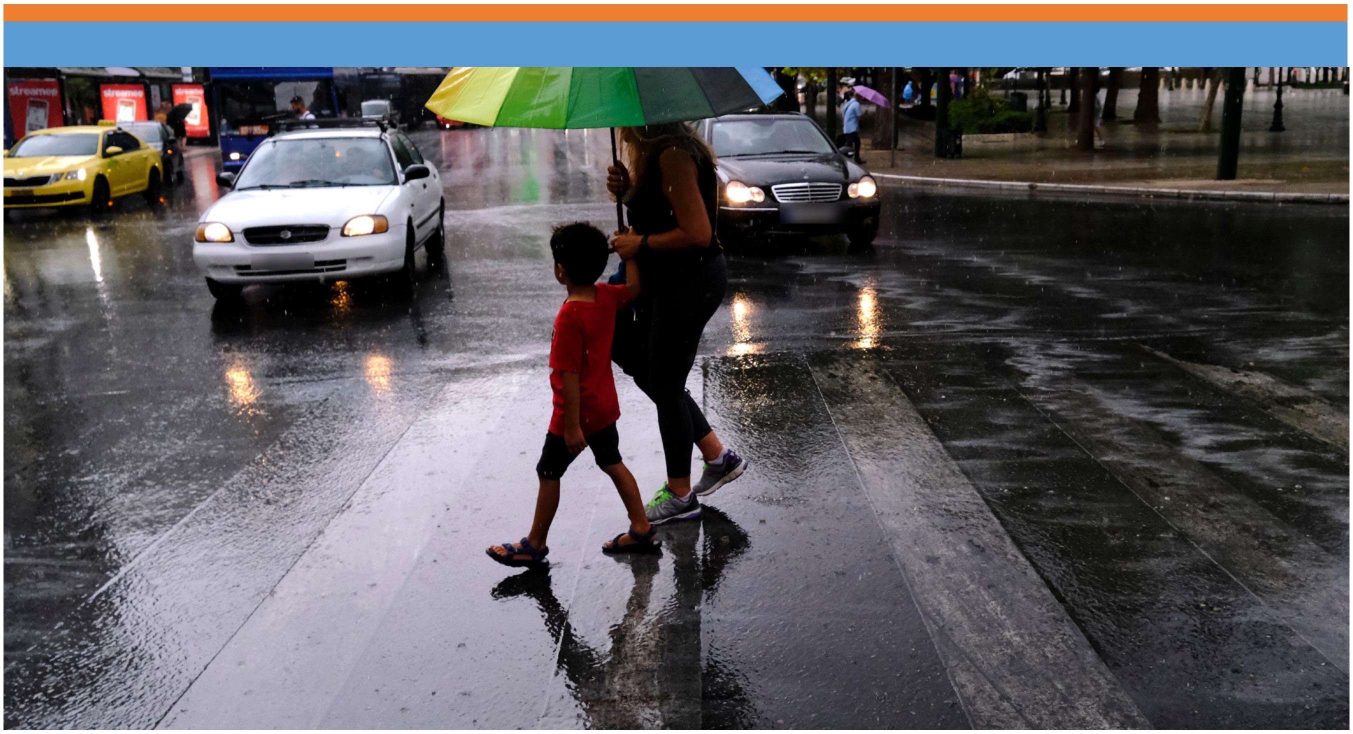
Other works have performed sun glare analysis in large metropolitan areas. In Mitra (30), pedestrian crashes were analyzed at 291 signalized intersections in Tucson, Arizona. Tucson's arterial street network is ideally situated for an analysis of sun glare because the grid pattern is aligned with the cardinal directions. Their reported analysis found an increase in the rate of glare-involved crashes during the morning peak period in March and the evening peak period in September, consistent with solar positioning during the peak commute periods in those months. The study also found that right-angle and rear-end crashes occurred at elevated rates during sun glare periods, but that crash severity generally did not increase due to sun glare.

Ma et al. (29) conducted a case-control study of pedestrian fatalities in Taiwan to examine the role of sun glare in pedestrian-vehicle crashes. They found that glare was associated with increased frequency and severity of crashes, especially involving older drivers. They also suggested using visibility aids to increase the conspicuity of pedestrians during twilight and nighttime.

An analysis found that an increase in the rate of glare-involved crashes during the morning peak period in March and the evening peak period in September, consistent with solar positioning during the peak commute periods in those months.

Work such as Billah et al. (32) analyzed pedestrian-motor vehicle crashes in San Antonio. They discovered that drivers were at fault for the majority of crashes, though risk of injury was much higher when the pedestrian was at fault. They also heat mapped crash sites to determine areas of high risk and offered recommendations for remediating risk factors in those areas.

Other variables may be included to provide more detailed analysis. In Li et al. (33), crash and weather data are augmented by the inclusion of street-level photography available online. This allowed the researchers to account for mitigating factors like shade provided by buildings or trees. They were able to combine these data sources to develop maps that show the duration of sun glare at intersections in Cambridge, Massachusetts during sunrise and sunset time periods on dates of interest. Due to the size of the collected data, it was not feasible to include this information in the provided analysis.



Predictive analytics and artificial intelligence-based works such as Zhao et al. (34), Charm et al. (35), and Das et al. (36) use machine learning and data mining techniques to obtain actionable insights from CRIS data. These analyses predict the probability of crash occurrence at various intersections in Texas. One of these, Zhao et al. (34), used police records of pedestrian crashes to derive crash frequency models using machine learning algorithms. Their models indicated the influence of higher speed and intoxication of drivers or pedestrians on the severity of injuries. Work such as Omranian et al. (37) studied the impact of adverse weather conditions such as precipitation, on motor vehicle crashes. Their study included motor vehicle crashes for all of Texas for the year 2015 and evaluated the risk of crashes due to rainfall. The study found that rainfall increases the risk of accidents by 57% across the state of Texas. Omranian et al. used radar-imagery derived precipitation estimations to approximate rainfall events at a crash site, which introduces several limitations. Rainfall events near the radar sensor attenuate some of the energy, reducing the accuracy farther from the sensor. Additionally, the curvature of the Earth prevents accurate sensing of ground conditions. As a result, the researchers decided to use more established methods of data collection, including ground-based hourly weather measurement sites. These data would be interpolated to estimate prevailing weather conditions at crash sites.

A study discovered that drivers were at fault for the majority of crashes, though risk of injury was much higher when the pedestrian was at fault.

A study found that rainfall increases the risk of accidents by 57% across the state of Texas.

METHODOLOGY

Compiled Dataset

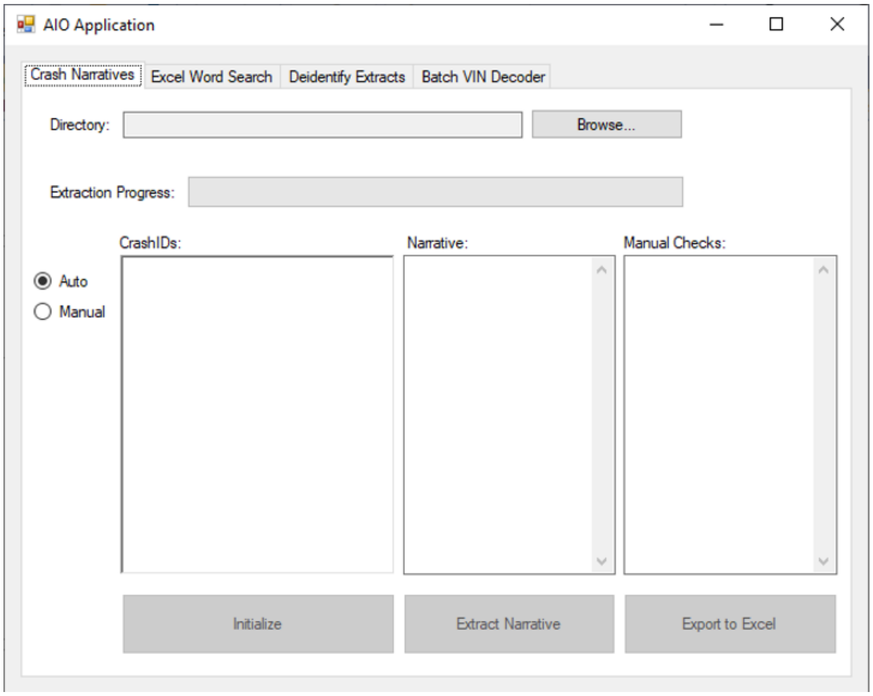
Researchers compiled 2018-2020 pedestrian crash data from TxDOT's crash record information system (CRIS) for five major Texas cities: Austin, Dallas, Fort Worth, Houston, San Antonio. The team selected the three major cities with the highest KAB/100k population: Austin, Dallas, and San Antonio as shown in Table 2. Note that Houston had more KABCO pedestrian crashes per capita than San Antonio but fewer KAB pedestrian crashes per capita.

Table 2. Pedestrian Crashes by City.

City	2018-2020					3-year Average	
	KABCO	KAB	2020 Pop.	KABCO/ 100k Pop.	KAB/ 100k Pop.	KABCO/ 100k Pop.	KAB/ 100k Pop.
Austin	1052	759	961,855	109.37	78.91	36.46	26.30
Dallas	1644	1112	1,304,379	126.04	85.25	42.01	28.42
Ft. Worth	671	424	918,915	73.02	46.14	24.34	15.38
Houston	2712	1553	2,304,580	117.68	67.39	39.23	22.46
San Antonio	1744	1132	1,434,625	121.56	78.91	40.52	26.30
Total	7,823	4,980	6,924,354	112.98	71.92	37.66	23.97

Researchers then developed a process to extract the desired 'unstructured' information from the crash reports. Using an in-house tool (Figure 4), researchers extracted the crash narratives and crash diagrams of all pedestrian related crashes from the three selected cities. Researchers removed personal identifiable information (PII) from the narratives such as VIN, license numbers, names, etc.

Figure 4. TTI's Crash Extraction Tool.



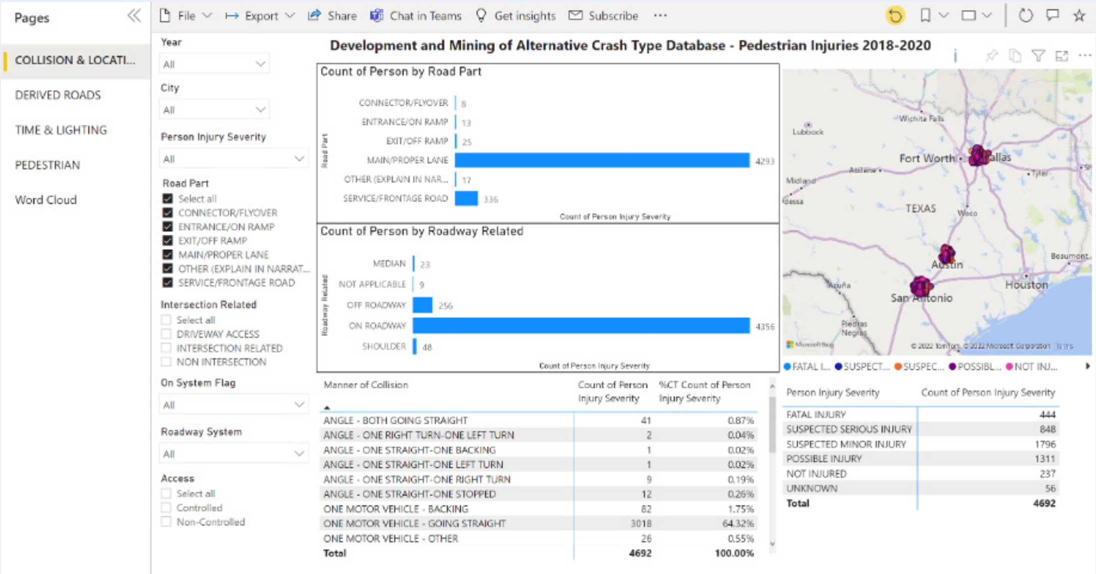
● Auto
○ Manual

The process involved using three Visual Studio applications, developed by TTI researchers, which automate the downloading PDF forms of the crash report, extracting the narrative section and the diagram section of the reports, respectively. The researchers did not use the PDF versions of the crash reports. Instead, they extracted the narrative data to multiple Excel files and manually reviewed the narratives to remove PII. The diagrams were stored as individual bitmap images. Any work involving PII information was done by TTI researchers authorized to access all CRIS data, including PII, under TTI Data Clearinghouse's Institutional Review Board (IRB) agreement. In addition to the crash report data obtained from the PDFs, CRIS data accessed on 10/7/2021, was also used. The narrative data, diagrams and CRIS data can all be linked based on the crash ID.

Developed Dashboard

It was important to see and analyze the crash data spatially. Thus, researchers created an online dashboard to visualize the data as shown in Figure 5. The dashboard can help provide data insights not easily seen in a traditional database/table. It is divided into five main pages: collision & location, derived roads, time & lighting, pedestrian, and word cloud. Most of the pages can be further filtered for a deeper dive into the data such as by year, city, injury severity, etc. Using the dashboard, an “Access” filter was created to differentiate between crashes on controlled access (freeways) or non-controlled access (non-freeways) facilities – which includes frontage road crashes. This filtering was done because of the different operating and crash characteristics on those facilities which greatly influence the types of countermeasures to be considered.

This dashboard is available to help practitioners conduct site investigations.



Development and Mining of Alternative Crash Type Database - Pedestrian Injuries 2018-2020

Year: All

City: All

Person Injury Severity: All

Road Part
- ☑ Select all
- ☑ CONNECTOR/FLYOVER
- ☑ ENTRANCE/ON RAMP
- ☑ EXIT/OFF RAMP
- ☑ MAIN/PROPER LANE
- ☑ OTHER (EXPLAIN IN NARRAT...
- ☑ SERVICE/FRONTAGE ROAD

Intersection Related
- □ Select all
- □ DRIVEWAY ACCESS
- □ INTERSECTION RELATED
- □ NON INTERSECTION

On System Flag: All

Roadway System: All

Access
- □ Select all
- □ Controlled
- □ Non-Controlled

Manner of Collision	Count of Person Injury Severity	%CT Count of Person Injury Severity
ANGLE - BOTH GOING STRAIGHT	41	0.87%
ANGLE - ONE RIGHT TURN-ONE LEFT TURN	2	0.04%
ANGLE - ONE STRAIGHT-ONE BACKING	1	0.02%
ANGLE - ONE STRAIGHT-ONE LEFT TURN	1	0.02%
ANGLE - ONE STRAIGHT-ONE RIGHT TURN	9	0.19%
ANGLE - ONE STRAIGHT-ONE STOPPED	12	0.26%
ONE MOTOR VEHICLE - BACKING	82	1.75%
ONE MOTOR VEHICLE - GOING STRAIGHT	3018	64.32%
ONE MOTOR VEHICLE - OTHER	26	0.55%
Total	**4692**	**100.00%**

Person Injury Severity	Count of Person Injury Severity
FATAL INJURY	444
SUSPECTED SERIOUS INJURY	848
SUSPECTED MINOR INJURY	1796
POSSIBLE INJURY	1311
NOT INJURED	237
UNKNOWN	56
Total	**4692**

Figure 5. Major Texas City Pedestrian Crashes Dashboard.

Descriptive Statistics

The crash dataset included 4442 pedestrian crashes across the three study cities. Table 3 shows the distribution of the crashes by number of cars and number of pedestrians. Most of the crashes (about 86%) involved one car and one pedestrian. About 96% of the crashes had no more than 2 cars and no more than 2 pedestrians. Table 4 shows the distribution of the crashes by severity (KABCO scale) and collision type. As expected, very few of the crashes are property damage only (PDO). In almost two thirds of the crashes, there was one straight-proceeding vehicle involved.

Table 3. Crash Distribution by Number of Units.

NUMBER OF CARS	PERCENT OF CRASHES BY NUMBER OF PEDESTRIANS					3-YEAR AVERAGE	
	1	2	3	4	5	6	Total
1	85.84	3.17	0.29	0.07	0.02	0	89.4
2	6.75	0.54	0.11	0.02	0	0.02	7.45
3	1.89	0.27	0.07	0.05	0	0	2.27
4	0.47	0.07	0	0	0	0	0.54
5	0.14	0	0	0	0	0	0.14
6	0.09	0	0	0	0	0	0.09
7	0.02	0.02	0	0	0	0	0.05
8	0.07	0	0	0	0	0	0.07
Total	**95.27**	**4.07**	**0.47**	**0.14**	**0.02**	**0.02**	**100**

Table 4. Crash Distribution by Severity and Collision Type.

CRASH SEVERITY	PERCENT OF CRASHES BY COLLISION TYPE				
	One vehicle, going straight	One vehicle, turning left	One vehicle, turning right	All other collision types	Total
K	9.12	0.14	0.16	0.63	10.04
A	14.36	2.32	0.74	1.31	18.73
B	23.37	9.64	3.78	2.12	38.9
C	15.62	7.18	3.35	1.78	27.94
PDO	2.25	0.79	0.41	0.77	4.21
Unknown	0.16	0.02	0	0	0.18
Total	**64.88**	**20.08**	**8.44**	**6.6**	**100**

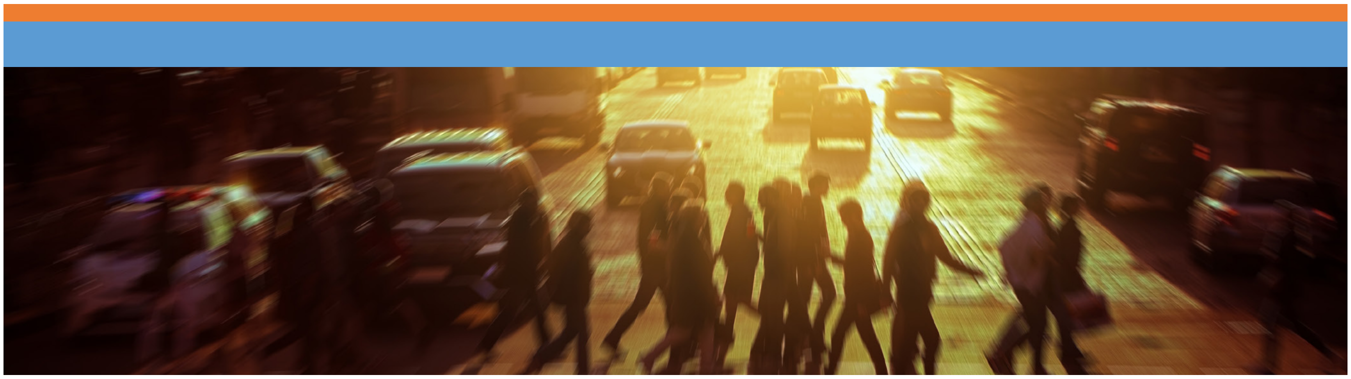


Table 5 shows the distribution of the crashes by year. The year 2020 shows a slight reduction in crashes compared to 2018 and 2019, perhaps due to COVID-related restrictions. Figure 6 shows the distribution of crashes by severity and year. The year 2020 shows a slight shift toward more severe crashes compared to 2018 and 2019.

Table 5. Crash Distribution by Year.

Year	Count	Percent of Crashes
2018	1537	34.60
2019	1664	37.46
2020	1241	27.94

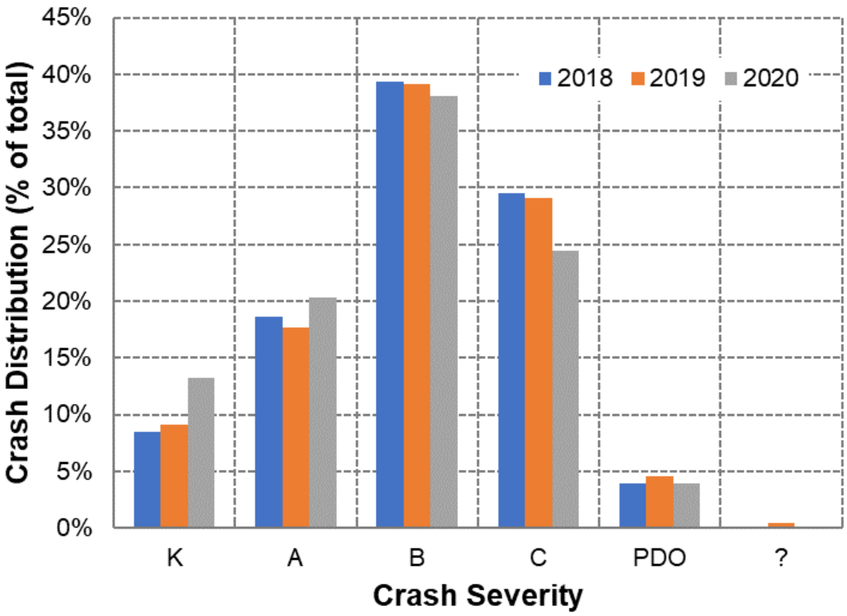


Figure 6. Distribution of Crashes by Severity and Year.

Table 6 and Table 7 shows the distribution of unit types by non-fatal injury and fatality count, respectively. Out of the 5124 vehicles involved in the crashes, only 274 had one or more injured occupants, and only 6 had a deceased occupant. Conversely, 3955 of the 4692 involved pedestrians were non-fatally injured and 444 were deceased. These distributions are intuitive for pedestrian-vehicle crashes.

Table 6. Crash Unit Distribution by Unit Type and Non-Fatal Injury Count.

TOTAL INJURY COUNT	UNIT TYPE			
	Car	Pedestrian	Other	Total
0	4850	737	12	**5599**
1	219	3955	1	**4175**
2	45	0	0	**45**
3	9	0	0	**9**
4	0	0	0	**0**
5	1	0	0	**1**
Total	**5124**	**4692**	**13**	**9829**

Table 7. Crash Unit Distribution by Unit Type and Fatality Count.

TOTAL FATALITY COUNT	UNIT TYPE			
	Car	Pedestrian	Other	Total
0	5118	4248	13	**9379**
1	6	444	0	**450**
Total	**5124**	**4692**	**13**	**9829**

Light and Weather Conditions

Figure 7 shows the distribution of crashes by the light condition variable in CRIS. This distribution roughly matches the distribution of light conditions in a typical day. Figure 8 shows the distribution of crashes by the weather condition variable in CRIS.

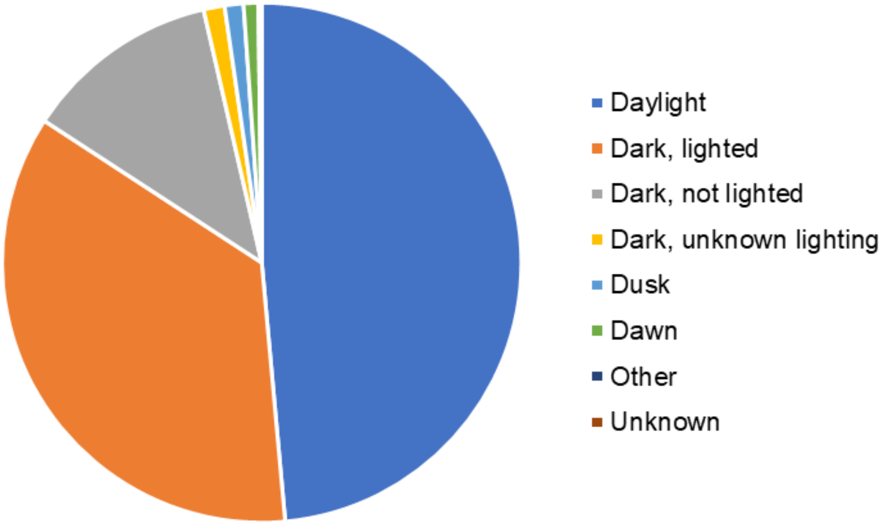


Figure 7. Distribution of Crashes by Light Condition.

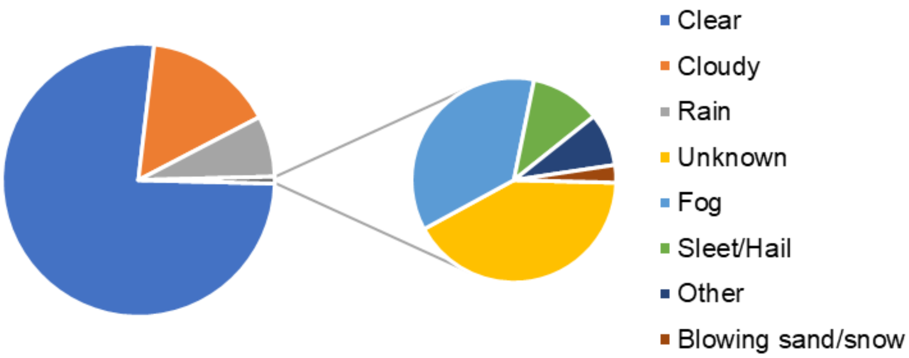


Figure 8. Distribution of Crashes by Weather Condition in CRIS Database.

Figure 9 shows the distribution of crashes by pavement condition. Note that the proportion of crashes occurring on wet pavement (coded as wet or standing water) is similar to the proportion of crashes occurring during rain in Figure 8.

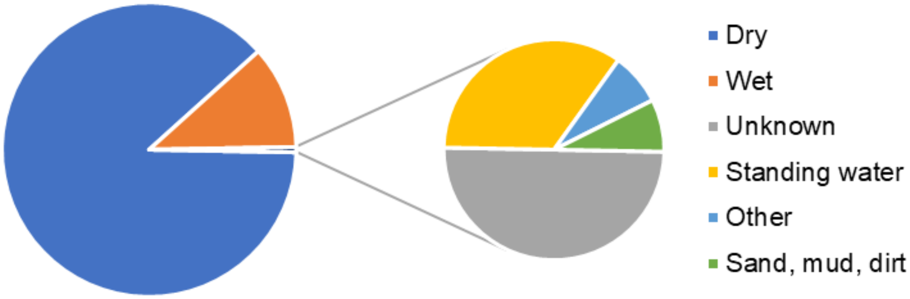


Figure 9. Distribution of Crashes by Pavement Surface Condition in CRIS Database.

Geography

Table 8 shows the distribution of the crashes by relation to intersection and collision type. The trends between these two variables are intuitive. For instance, most crashes involving a turning vehicle occur at an intersection or a driveway, and most crashes involving a through vehicle do not occur at an intersection or a driveway. Most of the crashes (93.4% of the total) are coded with a collision type that involves one vehicle proceeding straight or turning. The remainder of the crashes (6.6%) are either multiple-vehicle crashes or single-vehicle crashes involving a vehicle that was backing or performing a maneuver described as "other". Note, the total one-vehicle collision type of 93.4% in Table 8 is higher than the 89.4% in Table 3 because the numbers were obtained by querying different CRIS variables.

Table 8. Crash Distribution by Relation to Intersection and Collision Type.

RELATION TO INTERSECTION	PERCENT OF CRASHES BY COLLISION TYPE				
	One vehicle, going straight	One vehicle, turning left	One vehicle, turning right	All other collision types	Total
Non-intersection	45.27	0.43	0.05	3.65	**49.39**
Intersection-related	14.59	15.89	6.8	0.68	**37.96**
Intersection	4.12	1.69	0.43	1.53	**7.77**
Driveway access	0.9	2.07	1.17	0.74	**4.89**
Total	**64.88**	**20.08**	**8.44**	**6.6**	**100**

Table 9 shows the distribution of the crashes by road alignment and road class. The distributions across the two variables reflect the characteristics of the roadway network in the study cities. A few of the crashes occurred on roadways where pedestrians are prohibited. Table 10 shows the distribution of the crashes by access control. More than half of the crashes on interstates were on access-controlled roadways (i.e., the mainline, not the frontage roads), and almost all the crashes on tollways were on controlled-access roadways. Across the entire crash sample, 9.14% of the crashes occurred on controlled-access roadways.

Table 9. Crash Distribution by Road Alignment and Road Class.

ROAD ALIGNMENT	PERCENT OF CRASHES BY ROAD CLASS							
	City street	US & state highways	Interstate	FM	Other	Tollway	County road	Total
Straight, level	68.08	8.82	7.07	1.31	0.32	0.23	0.07	**85.88**
Straight, grade	5.72	1.1	1.08	0.07	0.02	0.02	0	**8.01**
Curve, level	1.6	0.32	0.52	0.02	0	0.07	0	**2.52**
Straight, hill-crest	1.33	0.18	0.34	0	0.05	0	0	**1.89**
Curve, grade	0.61	0.14	0.18	0.02	0	0	0	**0.95**
Curve, hillcrest	0.27	0.07	0.09	0	0	0	0	**0.43**
Other	0.16	0	0.02	0	0	0	0	**0.18**
Unknown	0.11	0	0.02	0	0	0	0	**0.14**
Total	77.87	10.63	9.32	1.42	0.38	0.32	0.07	**100**

Table 10. Crash Distribution by Access Control and Road Class.

ROAD CLASS	COUNT	PERCENT ON CONTROLLED-ACCESS ROADWAYS
City street	3459	0.20
US & state highways	472*	33.05
Interstate	414*	55.56
FM	63	0.00
Other	17	0.00
Tollway	14	92.86
County road	3	0.00

**Note: includes mainline and frontage/service road crashes.*

Time of Day

The following graphs show temporal distributions of the crashes:

- Figure 10. Distribution of Crashes by Month.
- Figure 11. Distribution of Crashes by Day of Week.
- Figure 12. Distribution of Crashes by Hour and City.

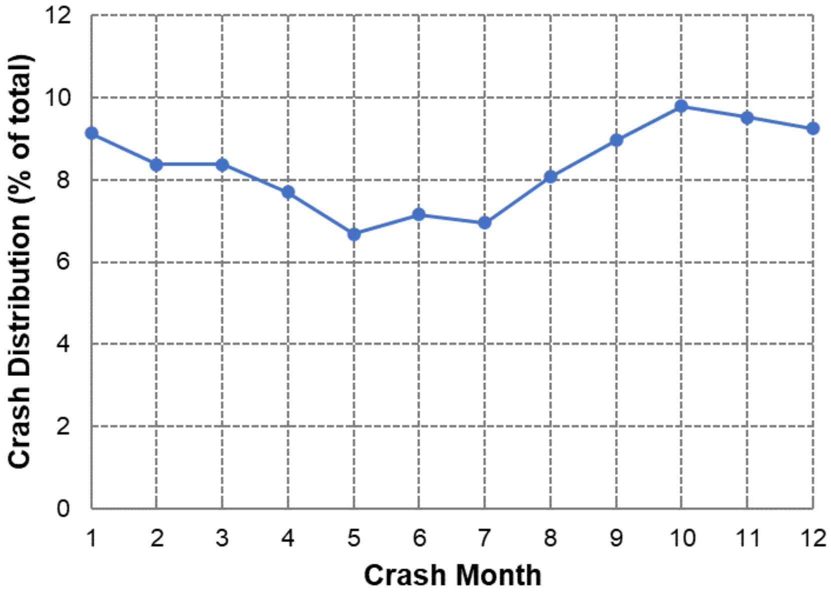


Figure 10. Distribution of Crashes by Month.

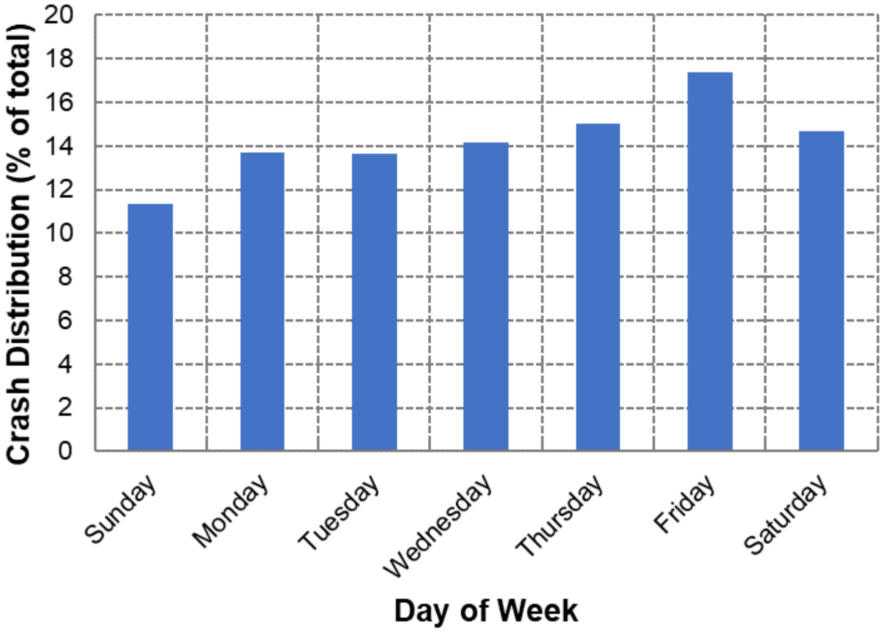


Figure 11. Distribution of Crashes by Day of Week.

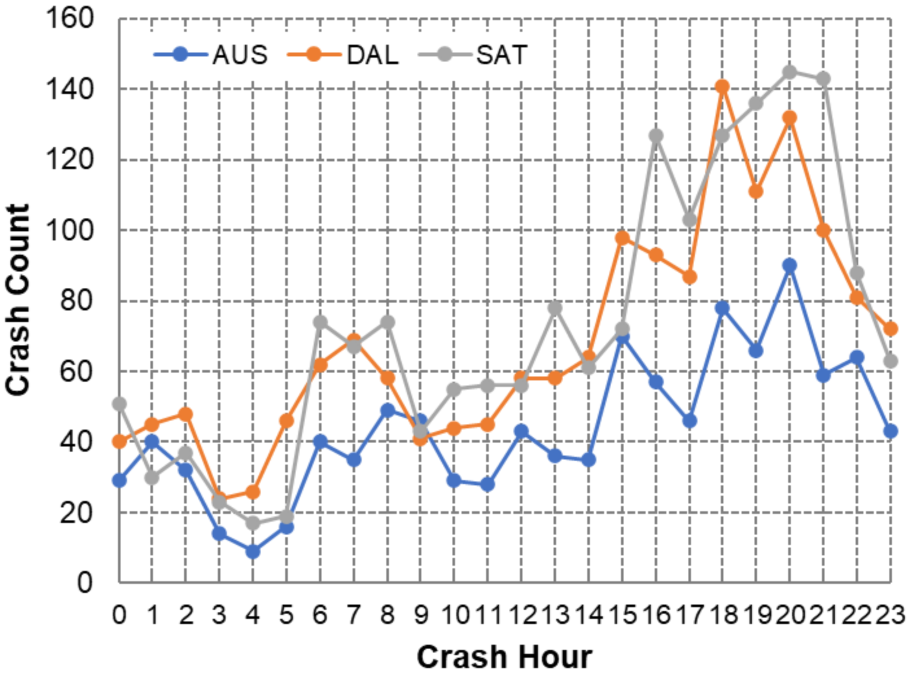


Figure 12. Distribution of Crashes by Hour and City.

Pedestrian-vehicle crashes temporally correlate with expected pedestrian volumes by time of day.

The trends in Figure 12 show that all three cities tend to have similar hourly trends. Pedestrian-vehicle crashes have their largest peak in the afternoon and evening hours of 3-11 PM, with smaller peaks in the morning hours of 6-8 AM and the lunch hour of 1 PM. This trend roughly reflects pedestrian volumes, but with somewhat more crashes in the evening hours, perhaps due to the presence of more pedestrians unfamiliar with the area (e.g., people going to dinner or bars). However, Austin generally experiences fewer crashes than the other cities in most hours.

Developed Training Dataset

The latest edition of PBCAT (v3) released in June 2001, classifies non-motorized crashes into 81 possible crash types between motorist and non-motorist, as shown in Table 11. One of the main goals of this project is to develop algorithms that can automatically determine the crash types based on the maneuvers (motorist and non-motorist) derived from the crash narratives and diagrams. For example, a simple illustration of a straight-crossing from right crash type (S-CR) is shown in Figure 13.

Non-Motorist Maneuver ***Motorist Maneuver***	**CR:** Crossing Path from Motorist's Right	**CL:** Crossing Path from Motorist's Left	**CU:** Crossing Path, Unknown Direction	**PS:** Parallel Path Same Direction	**PO:** Parallel Path Opposite Direction	**PU:** Parallel Path Unknown Direction	**MU:** Moving in Unknown Path/ Direction	**ST:** Stationary	**OU:** Other/ Unusual	**UN:** Unknown	**FC:** Non-motorist Fall or Crash
S: Going Straight	S-CR	S-CL	S-CU	S-PS	S-PO	S-PU	S-MU	S-ST	S-OU	S-UN	n/a[2]
R: Turning Right	R-CR	R-CL	R-CU	R-PS	R-PO	R-PU	R-MU	R-ST	R-OU	R-UN	
L: Turning Left	L-CR	L-CL	L-CU	L-PS	L-PO	L-PU	L-MU	L-ST	L-OU	L-UN	
P: Parked	P-CR	P-CL	P-CU	P-PS	P-PO	P-PU	P-MU	n/a	P-OU	P-UN	
E: Entering Traffic Lane	E-CR	E-CL	E-CU	E-PS	E-PO	E-PU	E-MU	E-ST	E-OU	E-UN	
B: Backing	B-CR	B-CL	B-CU	B-PS	B-PO	B-PU	B-MU	B-ST	B-OU	B-UN	
O: Other Maneuver	O-CR	O-CL	O-CU	O-PS	O-PO	O-PU	O-MU	O-ST	O-OU	O-UN	
U: Unknown Maneuver	U-CR	U-CL	U-CU	U-PS	U-PO	U-PU	U-MU	U-ST	U-OU	U-UN	
N: Non-Collision	Not Applicable – No Crash type returned										N-FC

Table 11. Crash Type Matrix detailed from Motorist and Non-Motorist Maneuver Selections (10).

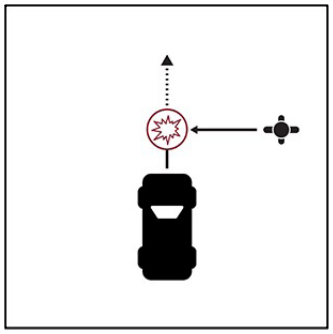


Figure 13. Straight-Crossing from Right Crash Type (4).

Another goal of the project is to determine the “intentionality” of the pedestrian involved. Researchers developed a process to determine whether the non-motorist was an ‘intended’ or ‘unintended’ pedestrian. This was done by reviewing crash narratives and inferring why the pedestrian was at the crash scene (i.e., what was the circumstance). Generally, the pedestrian was deemed ‘unintended’ if they were associated with a vehicle or other mode (e.g., getting out of a vehicle and being struck) as shown in examples in Table 12. Otherwise, they were considered as an ‘intended’ pedestrian because they were likely at the scene for walking purposes. Note, the intentionality is not always possible to determine due to insufficient information.

Researchers developed a process to determine whether the non-motorist was an ‘intended’ or ‘unintended’ pedestrian. This was done by reviewing crash narratives and inferring why the pedestrian was at the crash scene.

Table 12. Pedestrian Intentionality Matrix.

EXAMPLE CIRCUMSTANCES	INTENTIONALITY
Crossing roadway, crossing street	Intended (unless if associated with getting out of a vehicle)
Fleeing police	
Jumping from bridge	
Jumping from car	
Standing in traffic	
Standing on median, shoulder, or off the road	
Suicide	
Walking along the sidewalk	
Walking or lying down in traffic	
Walking or lying down on median, shoulder, or off the road	
Previous crash	Unintended
Stalled or stopped vehicle	
Working	
Sitting at bus stop	

Researchers created a representative training dataset to develop the algorithms necessary to classify the crash types and pedestrian intentionality. This dataset was used to train the various models tested. The training dataset consisted of 1000 random crashes (equally by city) and access type (Table 13) were reviewed and classified manually for crash type (Table 14) and intentionality (Table 15). Appendix A shows examples of the crash review process. Note that the crash types could not be determined for some crashes.

Table 13. Training Dataset by City and Access Type.

ACCESS	AUSTIN	DALLAS	SAN ANTONIO	TOTAL	PERCENT OF TOTAL
Controlled (Fwy.)	39	81	27	147	15%
Non-Controlled (Non-Fwy.)	294	253	306	853	85%
Total	333	334	333	1000	100%
Percent of Total	33%	33%	33%	100%	100%

Table 14. Training Dataset by Crash Type.

VEH/PED	CR	CL	CU	PS	PO	PU	MU	ST	OU	UN	FC	TOTAL
S	209	215	99	46	11	20	9	81	11	30	0	**731**
R	29	7	4	15	5	5	0	3	0	0	0	**68**
L	5	6	6	29	45	18	0	4	1	2	0	**116**
P	0	0	1	0	0	0	0	0	2	0	0	**3**
E	3	3	3	0	0	0	0	4	0	0	0	**13**
B	1	1	1	0	0	0	1	10	4	1	0	**19**
O	1	2	2	1	1	4	0	17	2	2	0	**32**
U	0	0	2	0	0	0	2	2	0	2	0	**8**
N	0	0	0	0	0	0	0	0	0	0	8	**8**
Total	**248**	**234**	**118**	**91**	**62**	**47**	**12**	**121**	**20**	**37**	**8**	**998**

Note: two crashes did not involve a pedestrian.

Table 15. Training Dataset by Pedestrian Intentionality.

INTENTIONALITY	AUSTIN	DALLAS	SAN ANTONIO	TOTAL	PERCENT OF TOTAL
Intended	201	214	287	702	70%
Unintended	131	117	39	287	29%
NA	1	3	7	11	1%
Total	**333**	**334**	**333**	**1000**	**100%**

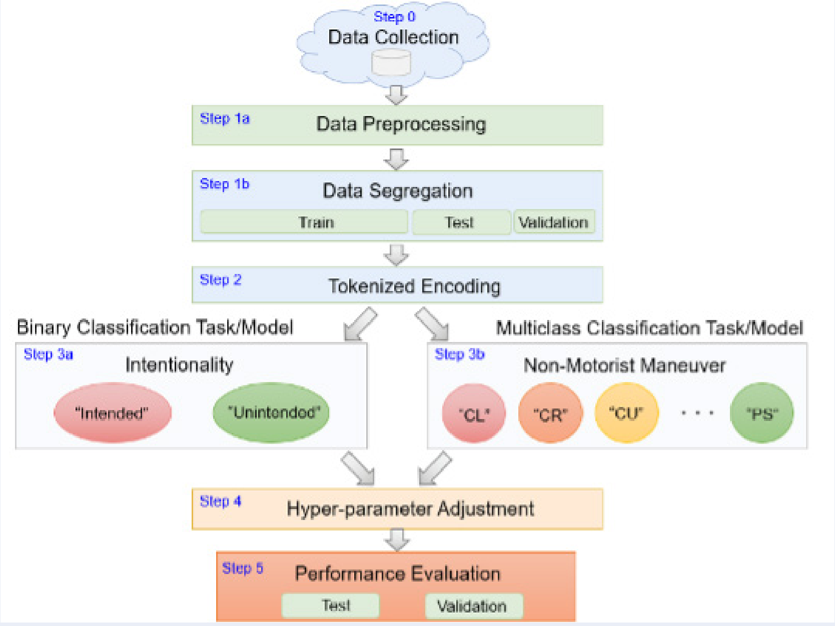
Note: intentionality could not be determined for 11 crashes (NA).

PBCAT Crash Typing

The problem of dataset imbalance is frequently encountered in crash narrative classification. For example, much of the pedestrian's maneuver in the training dataset is either a crossing or parallel type as shown in Table 14. Likewise, most of the vehicle's maneuver is straight or right/left turns. Thus, researchers undertook this classification task while considering the increasing utilization of large language models (LLM)s and the growing accessibility of Natural Language Processing (NLP) as a valuable tool for safety research. This effort is described as a multiclass classification task focusing on the pedestrian maneuver/action. The transfer learning approach was employed, utilizing the pre-trained BERT (38) and Robust Optimized BERT Approach (RoBERTa) (39) models. The Transformers open-source library (40) and the ktrain low code Python library (41) for deep learning facilitated the development of the pedestrian maneuver classification framework (Figure 14). After data preparation and initial preprocessing, the narrative data was converted into embeddings suitable for each model using the encode function of the Transformer library's tokenizer. Each pre-trained model was then expanded by incorporating a linear transformation layer representing the one-hot-encoded representation of non-motorist maneuver classes (CL: Crossing Path from Motorist's Left, CR: Crossing Path from Motorist's Right, CU: Crossing Path, Unknown Direction, PO: Parallel Path Opposite Direction, ST: Stationary, PU: Parallel Path Unknown Direction, PS: Parallel Path Same Direction, and OT: Other). Hyperparameters were adjusted based on learning rate range tests (42). Subsequently, the models underwent training, validation, and testing, with the final evaluation of performance conducted based on the test and validation subsets.

The problem of dataset imbalance is frequently encountered in crash narrative classification.

Figure 14. PBCAT crash typing framework



Narrative Data Preprocessing

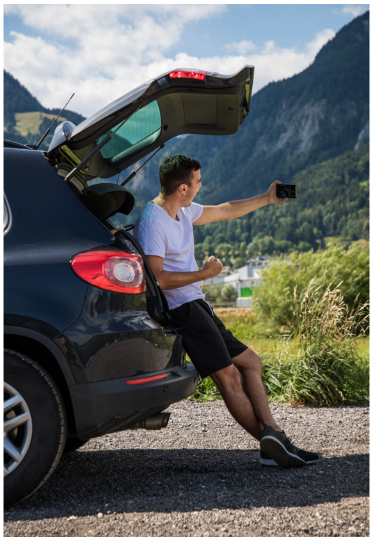
An early analysis of the distribution of classes revealed that some of the non-motorist maneuver classes represented a small portion of the 1,000-training dataset, e.g., “Unknown” (UN) with 37 samples, “Moving in Unknown Path/Direction” (MU) with 12 samples, “Non-motorist Fall or Crash” (NA) with 6 samples, “Other/Unusual” (OU) with 20 samples. In the context of this research, the resulting classification is a “Detailed” classification. As a result, we combined these classes under the umbrella term “Other” (OT) in tests. While the PBCAT 3.0 has established a standardized protocol for crash type classification, enhancing data integration and streamlining the process, the imbalance in the resulting data can make it difficult to develop automated solutions, especially with smaller sample sizes. By merging smaller yet conceptually adjacent classes, the effects of dataset imbalance and size can be mitigated. So, researchers also tested a classification referred to in this paper as the “Abstract” classification containing “Crossing” (C), “Parallel” (P), “Stationary” (S), and “Other” (O) classes. Table 16 provides a detailed breakdown of the pedestrian maneuver data using PBCAT 3.0. Note that the crash types could not be determined for four crashes because two did not involve a pedestrian. For purposes of this study, these crashes were classified as “OT” and “O” in the Detailed and Abstract classifications, respectively.

Table 16. Training Dataset by Crash Type.

VEHICLE / PEDESTRIAN	CR	CL	CU	PS	PO	PU	ST	MU	OU	UN	FC	TOTAL
Detailed	CR	CL	CU	PS	PO	PU	ST	OT				
Abstract	C			P			S	O				
Straight	209	215	99	46	11	20	81	9	11	30	0	**731**
Right	29	7	4	15	5	5	3	0	0	0	0	**68**
Left	5	6	6	29	45	18	4	0	1	2	0	**116**
Parked	0	0	1	0	0	0	0	0	2	0	0	**3**
Entering	3	3	3	0	0	0	4	0	0	0	0	**13**
Backing	1	1	1	0	0	0	10	1	4	1	0	**19**
Other	1	2	2	1	1	4	17	0	2	2	0	**32**
Unknown	0	0	2	0	0	0	2	2	0	2	0	**8**
Non-collision	0	0	0	0	0	0	0	0	0	0	8	**8**
Total	**248**	**234**	**118**	**91**	**62**	**47**	**121**	**12**	**20**	**37**	**8**	**998**

Note: CR – Crossing path from motorist’s right; CL – Crossing path from motorist’s left; CU – Crossing path Unknown direction; PS – Parallel path Same direction; PO – Parallel path Opposite direction; PU – Parallel path Unknown direction; MU – Moving in Unknown path/direction; ST – Stationary; OU – Other/Unusual UN –Unknown; FC – Non-motorist Fall or Crash

Addressing the challenges specific to this task was crucial. Crash reports presented a unique narration style that posed difficulties for laypersons to comprehend and for an NLP model without prior exposure to such narration to be fine-tuned. Certain sections of text were often replaced with placeholders like "[retracted]," or sequences of symbols such as "*" or "#" frequently appeared, causing disruptions in the sentence flow. To improve text coherence, it was necessary to remove numerical information related to blood alcohol level, street addresses, dates and times, and case numbers. Additionally, the lack of specific references to pedestrians using terms like "pedestrian" or "non-motorist" contributed to text fragmentation, as parties involved in crashes were often referred to as "unit 1" or "unit #2", regardless of whether they were motorists or non-motorists. Furthermore, Non-ASCII (American Standard Code for Information Interchange) characters were also observed in the texts, necessitating appropriate handling to ensure accurate analysis.



Researchers adopted a minimalist approach to overcome the text cleaning challenges, considering the powerful nature of models like BERT and RoBERTa, which excel in a contextual understanding of words. Excessive data cleaning could potentially harm the performance of these models. Our approach involved removing Non-ASCII characters, redacted sections, and sequences of repeating characters (e.g., '#'). Additionally, we filtered out numbers longer than one character and repeated space characters. Empty parentheses and brackets that no longer contained text after cleaning were also filtered out. As a result, the narratives retained their original sentence structure. This approach ensures the optimal performance of BERT and RoBERTa by maintaining the contextual integrity of the text.

To fine-tune the three NLP models for motorist maneuver classification, the researchers partitioned the dataset into three sections: train, validation, and test. These sections were randomly sampled to ensure an equal distribution of classes. The train set, consisting of 70% of the samples, was utilized for fine-tuning the models. The validation data, accounting for 20% of the dataset, served to assess the models for overfitting, a phenomenon where models excel on familiar data but may suffer in general performance. Lastly, we employed the test set, comprising the remaining 10% of the dataset, to comprehensively evaluate various aspects of the model's overall performance.

Natural Language Processing Models

The introduction of the Transformer architecture by Vaswani et al. (43) revolutionized the field of NLP and led to significant improvements over previous state-of-the-art networks. However, before the advent of transformer-based models, achieving human-level language processing capabilities seemed beyond the reach of computers. In this study, researchers employed two transformer-based models sharing the same architecture to assess their impact on the classification quality of imbalanced datasets—a common challenge encountered in transportation safety research (14-16). These models process input text as tokens, which can be complete words, word segments, or characters, and convert them into embeddings through their initial layer. Each token is mapped to the model's internal vocabulary, represented as a sequence of numbers. The models come in two variants: BASE and LARGE, with input sequence length limits of 512 and 768 embeddings, respectively. As this dataset's embedding sequence length fell below the 512-embedding limits of the BASE version, researchers used the BASE variant. The researchers proceeded with the following two NLP models:

- **BERT:** BERT (38) is one of the most impactful natural language models today, accelerating NLP research and popularizing transfer learning in this field. It stands for Bidirectional Encoder Representations from Transformers. Unlike its preceding models, such as Word2Vec (44) and GloVe (45), which used word-level embeddings, BERT uses word pieces, making it more capable of handling out-of-vocabulary words. The model takes a bidirectional approach to language representation, meaning its comprehension of words, or more accurately said, word embeddings, is informed by the context of text following and preceding it. This model has been originally trained for two tasks. The mask language modeling task (46) involves masking a portion of text (represented as word embeddings) and training the model to predict the masked word. The next sentence prediction task involves training the model to comprehend the relationship between two sentences. It achieves this by learning to identify whether a given sentence is followed by another specific sentence.

- **RoBERTa:** The RoBERTa model (39) is a more Robustly Optimized version of BERT demonstrating a better or close performance to this model. Most of this model's improvement is derived from the modifications that Liu et al. (39) have made to BERT's training process. These modifications included training for a longer duration, using larger batch sizes, and utilizing a larger amount of training data (CC-NEWS dataset in addition to other English-language corpora). This model has been solely trained for the masked language modeling objective by applying a dynamically changing masking pattern to the training data in contrast to BERT, which has also been trained for next-sentence prediction. It is worth mentioning that RoBERTa's extensive vocabulary, which it has gained due to its larger training data, can help it better capture some linguistic nuances.

Multiclass Classification

In the initial data analysis, the researchers identified that the dataset was noticeably imbalanced. In a previous study on using BERT to classify crash severity types, researchers identified the dataset imbalance as an issue and recommended further exploration of augmentation methods and hyperparameter adjustments as possible avenues for improving the model's performance (14). Researchers tested two batch sizes of 32 and 64 to perform a fine-tuning simulation for five epochs in which they linearly increased the learning rate. This approach allowed the researchers to identify a suitable starting learning rate for fine-tuning each model. To reduce the effects of imbalance on the model's performance, researchers tested the Categorical Cross Entropy *(CCE)* (47), Balanced Categorical Cross Entropy *(BCE)*, and Focal Categorical Cross Entropy *(FCE)* (48) loss functions in fine-tuning the models. Researchers used the Adam optimizer (49) and a cyclical learning rate policy (42) to train the models. The loss values are calculated as follows:

$$CCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1_{y_i \in C_j} log_{p_{model}[y_i \in C_j]} \qquad (1)$$

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1_{y_i \in C_j}\, w_j\, log_{p_{model}[y_i \in C_j]} \qquad (2)$$

$$FCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} 1_{y_i \in C_j}\, \alpha \left(1 - p_{model}\left[y_i \in\ C_j\right]\right)^{\gamma} log_{p_{model}[y_i \in C_j]} \qquad (3)$$

where *N* and *C* are the number of samples and categories. $p_{model}\left[y_i \in\ C_j\right]$ represents the model's predicted probability of the sample of index *i* belonging to the class of index *j*. The w$_j$ represents the balancing weight that is set inverse to jth class's frequency. The focusing parameter is represented as *γ* (*α* and *γ* were set to 0.25 and 2 based on Lin et al.'s [48] experimental results).

In the assessment of models, the researchers considered several performance metrics. The Accuracy (ACC) represents the portion of the correct models' predictions. However, this metric does not discriminate between the models' predictions of the minority or the majority class. In datasets with noticeable imbalances, the prediction of the smaller classes affects this metric the least, even though models struggle the most to predict the underrepresented classes correctly. The researchers considered Precision as a measure of the accuracy of positive predictions and Recall as the measure of the model's ability to identify positive instances correctly. As a result, the macro averaged F1, the harmonic mean of the Precision and Recall, acts as the key performance indicator in this research.

The researchers considered Precision as a measure of the accuracy of positive predictions and Recall as the measure of the model's ability to identify positive instances correctly.

In the assessment of models, the researchers considered several performance metrics. The Accuracy (ACC) represents the portion of the correct models' predictions. However, this metric does not discriminate between the models' predictions of the minority or the majority class. In datasets with noticeable imbalances, the prediction of the smaller classes affects this metric the least, even though models struggle the most to predict the underrepresented classes correctly. The researchers considered Precision as a measure of the accuracy of positive predictions and Recall as the measure of the model's ability to identify positive instances correctly. As a result, the macro averaged F1, the harmonic mean of the Precision and Recall, acts as the key performance indicator in this research.

Precision (p): Precision measures positive patterns that are correctly predicted over the total positive prediction patterns. Precision is calculated for each class separately, so Macro Average Precision is calculated as a holistic measure of model's precision and is given as:

$$Average\ Precision = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}}{C} \tag{4}$$

Recall (r): Recall is a measure of positive patterns over the total correct predictions. Much like Precision, Recall is calculated for each class thus averaging is required for multiclass model assessment. This macro average is given as:

$$Average\ Recall = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + TN_i}}{C} \tag{5}$$

F-measure (F1): F-measure is the harmonic mean between recall and precision values. The macro averaged F-measure is calculated using the macro averaged recall and precision, respectively.

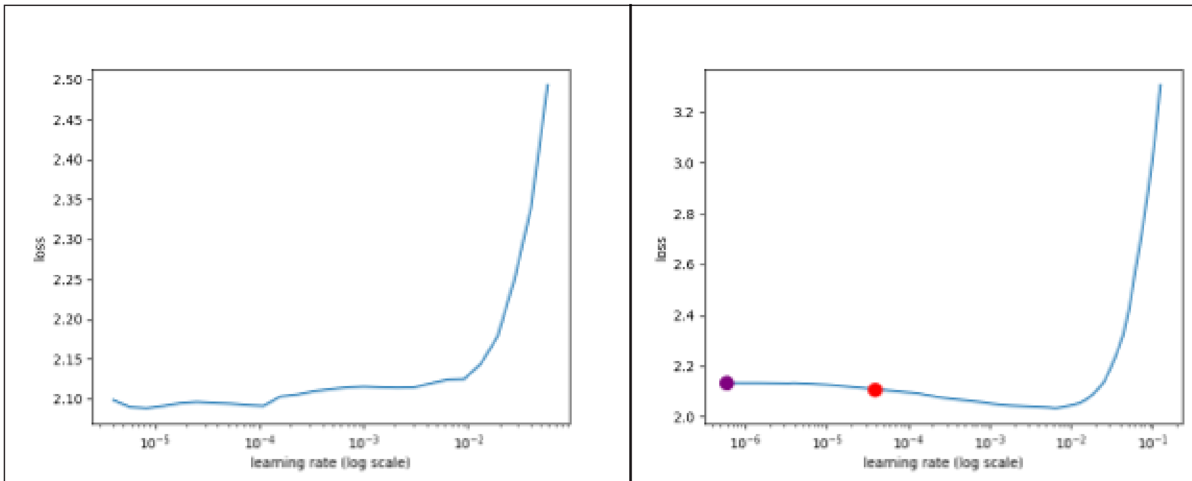
$$Average\ F1 \ = \sum_{i=1}^{C} \frac{2\,TP_i}{2\,TP_i \ + \ FP_i + \ FN_i} \tag{6}$$

Accuracy (acc): The prediction accuracy of a model in the context of classification is the ratio of correct predictions over the total number of examined instances. The Accuracy is given as:

$$Accuracy = \frac{(TP \ + \ TN)}{(TP + TN + FP + FN)} \tag{7}$$

where C represents the number of classes TP_i and TN_i are the correct predictions of sample belonging and not belonging to the i^{th} class. FP_i and FN_i represent the false predictions of sample belonging and not belonging to the i^{th} class.

The researchers started tests by performing a fine-tuning simulation for five epochs using 32 and 64 batch sizes in combination with several loss functions. These loss functions included Categorical Cross Entropy (CCE), which is one of the most widely used loss functions for multiclass tasks, and Balanced Categorical Cross Entropy (BCE), which applies weights to the loss value of each class inversely proportional to their frequency in the dataset, and Focal Categorical Cross Entropy (FCE), which has proven to be highly effective in encouraging the model to focus on the more challenging samples in the dataset (48). In the learning rate tests, some hyper-parameter combinations did not prove a noticeable decrease in loss values. Figure 15 demonstrates a hyper-parameter combination that did not prove a noticeable descent in loss (left side) and a combination that shows a point of steepest descent marked by a red dot and a point of minimum gradient marked by the purple dot (right side). As a result, the researchers selected a learning rate that fell in the shared range between these two points in hyper-parameter combinations that demonstrated these points.



model = RoBERTa, batch size = 64, loss = BCE

model = BERT, batch size = 32, loss = CCE

Figure 15. Examples of Fine-Tuning Simulation for Learning Rate Search.

The fine-tuning process is carried through iterations (epochs) in which model weights are adjusted to perform classification on the training data subset better. In the fine-tuning, the process was initiated with the pre-trained model parameters. The loss value was calculated for training data in each iteration, and the Adam optimizer was used to adjust the model parameters. At the end of each iteration, the loss was calculated for the validation data, and a snapshot of model parameters was stored alongside the loss and accuracy values. To prevent over-fitting, in which the model would specialize in classifying training data at the cost of its general performance, the researchers selected the model parameters from the epoch that demonstrated the lowest validation loss value. Figure 16 compares the loss values of the two best-performing models (RoBERTa) and best performing BERT model. This graph marks the point in the fine-tuning process that achieved the lowest validation loss. The researchers used the model parameters from this stage of the training for our evaluations. The corresponding model accuracies are shown in Figure 17 and demonstrate the less performant model's stagnant accuracy compared to the better-fine-tuned models.

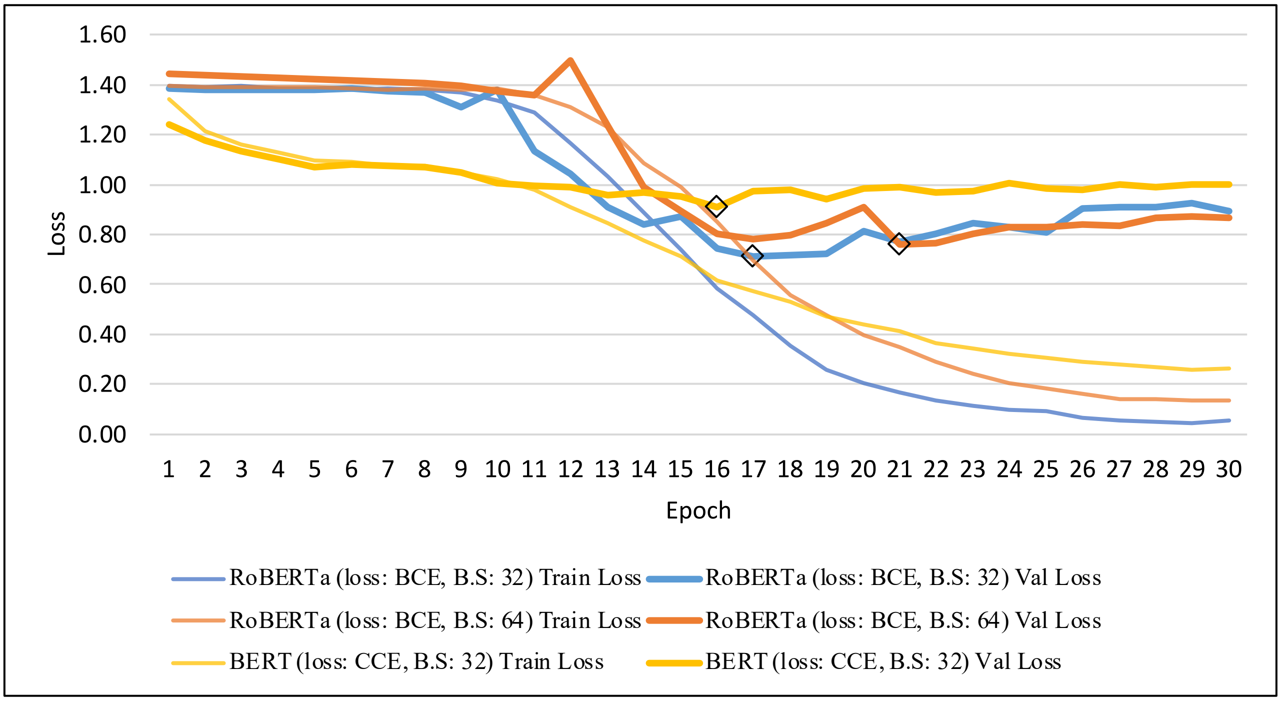


Figure 16. Comparison of Loss Values.

For the multiclass classification task, the researchers first tested the capability of models using eight classes. In these series of tests, the highest accuracy achieved was 34% and the highest average F1 was 25%. These tests did not demonstrate a significant advantage in using BCE loss as a strategy to improve the model performance. However, in the next step, the researchers performed the tests with a more abstract classification system comprised of four classes. In the tests using RoBERTa Base on the abstract classification system with four classes, the researchers observed the best performance which was 65.2% accuracy and 50.9% Average F1. The improvement in F1 performance was observed in the majority of the models using BCE. This finding indicated that multi classification with smaller number of classes can yield more accurate results (in terms of F1) and the BCE loss function can yield a more balanced result.

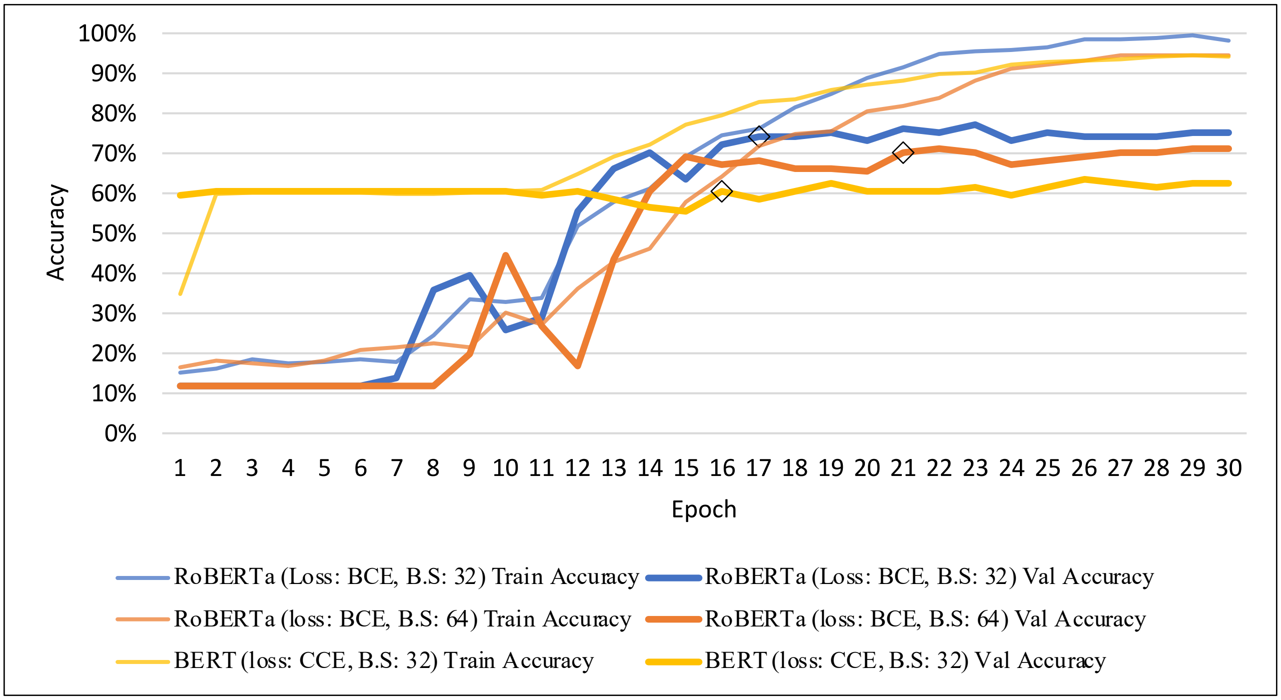


Figure 17. Comparison of Accuracies.

Combined Natural Language Processing and Computer Vision

Given the results and findings above, the researchers used a multi-step, multiclassification approach where the crashes were initially classified as either crossing or non-crossing. Figure 18 shows that 218 of the 600 (36%) crashes were correctly classified as crossing when the accuracy was 90% or more. Likewise, 109 of 400 (27%) crashes were correctly classified as non-crossing when the accuracy was 90% or more. Note that both classes had three outliers that were misclassified, which was deemed acceptable.

Researchers then explored CV and machine learning techniques to see if they can be applied to the crash diagrams to further subclassify the crash types i.e., pedestrian maneuvers.

Computer Vision and Machine Learning Exploration

Analogous to the crash narratives, the crash diagrams contain a substantial amount of data. However, crash diagrams seem to have more variability because there is a wider range of details that officers can depict in them. Given that there is no standardized process to create the diagrams, interpreting them even by humans, let alone computers/machines, can be a very challenging problem. However, researchers developed a three-step process to apply CV and machine learning to see if it is possible to infer pedestrian maneuvers from the crash diagrams (Figure 19).

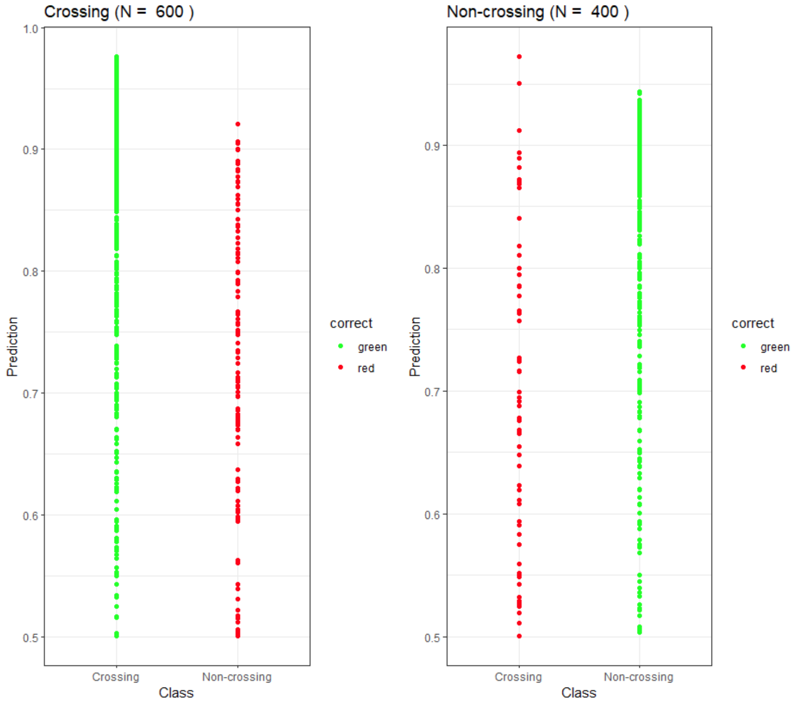


Figure 18. Observed vs. Predicted Crossing and Non-Crossing Crash Type.

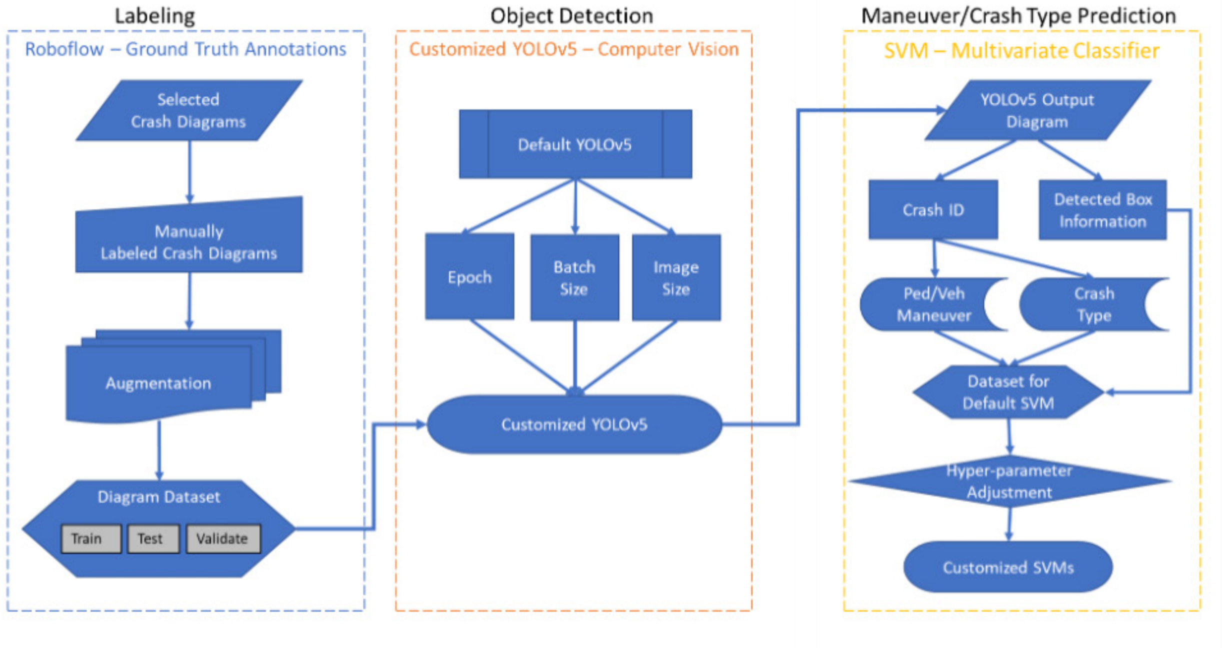


Figure 19. Interpreting Crash Diagram Framework.

Labeling

The first step was to select and manually annotate (label) a training dataset of diagrams. Researchers reviewed the crash diagrams within the 1000-crash training dataset and selected over 100 diagrams. The researchers labeled the following four elements (classes) that are required for each diagram. Other classes were also labeled but not subsequently used, such as the north arrow.

1. Initial Pedestrian Position.
2. Pedestrian Position at Point of Impact (POI).
3. Initial Vehicle Position.
4. Vehicle Position at POI.

Note that the initial and POI position could be indicated by an object and/or by a directional arrow. There were a number of diagrams that could not be used because they were missing these classes and/or the picture quality was poor.

Researchers used Roboflow (50) to label 62 diagrams for the test dataset. An augmentation process (51) was applied to add more varied examples to the training dataset. This is done to overcome "overfitting" because augmentation bridges the gap between the training dataset and the population. In this process, Roboflow generated more crash diagram to train by rotating them between -15 and +15 degrees clockwise and counterclockwise. Hence, the training dataset generated consisted of 112 training images, 12 validation images, and 12 test images (136 total).

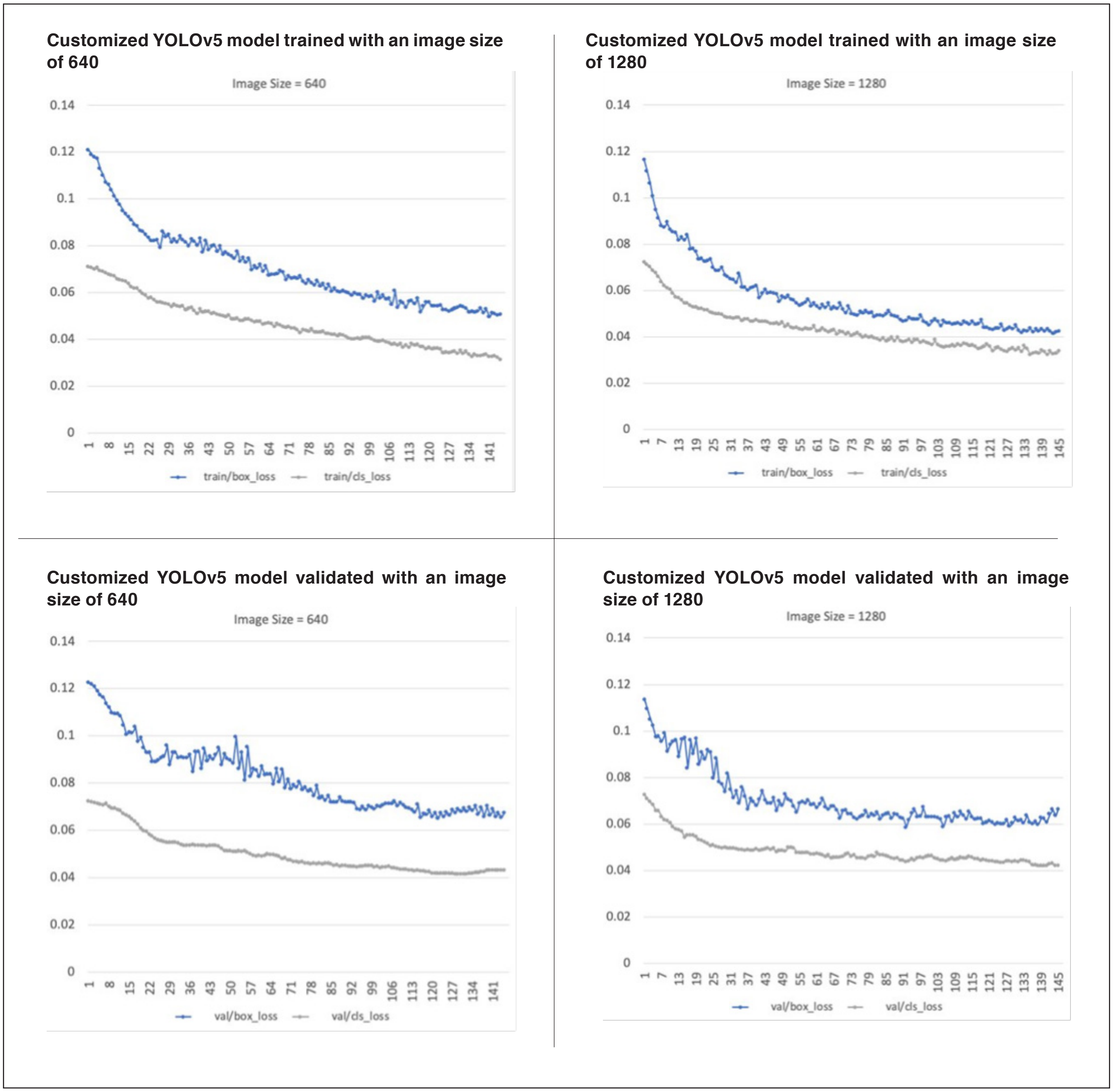
With a ground truth dataset generated by Roboflow, this study considered YOLOv5 – open-source and state-of-the-art CV method – to conduct object detection in a crash diagram. Only YOLOv5 was available at the start of this study. At that time, the choice was based on the performance and computational efficiency requirements of the study. For consistency, YOLOv5 continued to be used throughout the study.

Since there were not many objects per crash diagram, the study started with 100 epochs, and then increased to 300 epochs, and 1000 epochs. The customized YOLOv5 model eventually reached a point at 355 epochs, when accuracy did not change as number of epochs increased. This study also explored on changing the batch size from 16 to 32. Within this dataset, the accuracy result did not improve by enlarging the batch size, but the computation process time increased by minutes.

In addition to the common parameters (i.e., epoch and batch size), different values of image sizes were also compared in counter of detecting small object – pedestrian (*52*). Small objects are hard to identify because object detection models provide a prediction based on the loss function, which adds up the differences between prediction and ground truth in every pixel. When an object is small, it does not cover many pixels, then the signal is small and gets omitted when the detection model is trained. There are two practical techniques to effectively detect small objects:

1. Increase resolution of original images when establishing the ground-truth dataset.
2. Increase image size when inputting to an object detection model.

Because the original images were directly provided by the agent, this study implemented the second technique to achieve a better model. As shown in Figure 20, this study enlarged the default image size from 640 to 1280, and compared the “box_loss” – mean squared error (MSE) of the bounding box (the smaller the MSE the better the model), “cls_loss” – cross entropy loss of the box classification (the smaller the loss the better the model). At the same number of epochs (e.g., 100), although the image size did not reduce on the cross-entropy loss of the box classification, YOLOv5 with 1280 input image size had a smaller MSE than default YOLOv5.


Customized YOLOv5 model trained with an image size of 640

Customized YOLOv5 model trained with an image size of 1280

Customized YOLOv5 model validated with an image size of 640

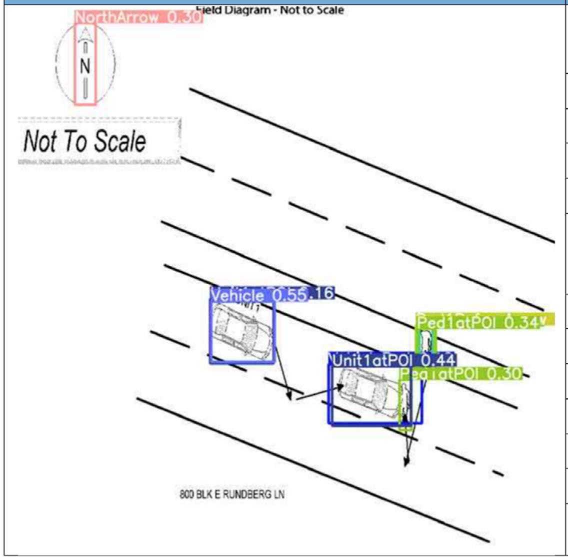
Customized YOLOv5 model validated with an image size of 1280

Figure 20. YOLOv5 Model Comparisons between Image Sizes (with First 145 Epochs).

Object Detection

A customized YOLOv5 model was performed under 300 epochs, with a batch size of 16, and an image size of 1280. The outputs of YOLOv5 were artificial intelligent (AI) labeled boxes and their labels. An example of an output crash diagram is as below. Because object detections did not lead to a classification of maneuvers or crash type, the relative positions of detected and identified objects would provide clues. A post-processed script was then written to output label information as shown in Table 17 (i.e., labels and coordinates of their left bottom corner of boxes – minimum of x-coordinate, minimum of y-coordinate, maximum of x-coordinate, maximum of y-coordinate).

Table 17. Examples of YOLOv5 Model Result Diagram and their Output Label Information.

YOLOv5 Result	Coordinates of Left Bottom Conner of Boxes				
	Label Index	*X* min.	*Y* min.	*X* max.	*Y* max.
 NorthArrow 0.30 Field Diagram - Not to Scale N Not To Scale Vehicle 0.55 16 Ped1atPOI 0.34 Unit1atPOI 0.44 Ped1atPOI 0.30 800 BLK E RUNDBERG LN	9	0.3119	0.5557	0.1021	0.1699
	3	0.5317	0.6228	0.3901	0.0390
	11	0.2992	0.5947	0.0782	0.1455
	2	0.8086	0.2463	0.0831	0.2780
	4	0.3457	0.6439	0.1942	0.0455
	5	0.3905	0.7907	0.0848	0.0350
	8	0.1811	0.1642	0.1251	0.2293
	5	0.6934	0.6130	0.0864	0.0341
	10	0.1366	0.8252	0.1185	0.1285
	10	0.2926	0.3732	0.0749	0.1236
	10	0.1465	0.1228	0.0807	0.1220
	10	0.1634	0.4354	0.0733	0.1179

Maneuver/Crash Type Prediction

With YOLOv5 detected and identified objects in boxes and their boxes' information, a supervised learning — Support Vector Machines (SVM), was used to classify the crash types. SVM is well-known (53, 54) for its ability to be universal approximators of any multivariate function to any desired degree of accuracy. Li et al. evaluated the SVM model for predicting motor vehicle crashes (55). They developed SVM and Negative Binomial regression models with data collected on rural frontage roads in Texas and compared SVM with NB regression and Back-Propagation Neural Network model from previous research. The result showed that SVM offers similar, if not better, performance than NB regression and BPNN models and overfits the data less. Li et al. suggested implementing SVM models when the purpose of the study is related to predicting motor vehicle crashes, especially with a sample size under 2000 observations.

The coordinates of the box corners are the input features for the SVM or any form of supervised learning classifier. Hence, classes (i.e., crash types, pedestrian maneuvers, vehicle maneuvers) were then post-processed by pulling the crash IDs from diagram name and matched to the features. Sample Data for SVM is shown in Table 18.

Table 18. Sample Data of Features and Classes.

CRASH ID	COORDINATES OF LEFT BOTTOM CORNER OF BOXES					CLASSES		
	Label Index	X minimum	Y minimum	X maximum	Y maximum	Crash Type	Ped Maneuver	Veh Maneuver
17524347	9	0.3119	0.5557	0.1021	0.1699	S-CL	CL	S
17524347	3	0.5317	0.6228	0.3901	0.0390	S-CL	CL	S
17524347	11	0.2992	0.5947	0.0782	0.1455	S-CL	CL	S
17524347	2	0.8086	0.2463	0.0831	0.2780	S-CL	CL	S
17524347	4	0.3457	0.6439	0.1942	0.0455	S-CL	CL	S
17524347	5	0.3905	0.7907	0.0848	0.0350	S-CL	CL	S
17524347	8	0.1811	0.1642	0.1251	0.2293	S-CL	CL	S
17524347	5	0.6934	0.6130	0.0864	0.0341	S-CL	CL	S
17524347	10	0.1366	0.8252	0.1185	0.1285	S-CL	CL	S
17524347	10	0.2926	0.3732	0.0749	0.1236	S-CL	CL	S
17524347	10	0.1465	0.1228	0.0807	0.1220	S-CL	CL	S
17524347	10	0.1634	0.4354	0.0733	0.1179	S-CL	CL	S

The concept of SVM is to map the data points into high dimensional space and find a hyperplane that can divide the points representing different classes. In this study, for pedestrian maneuvers, there are eight classes from the dataset: CR (crossing Right), CL (crossing left), CU (crossing unknown), PS (parallel same direction), PO (parallel opposite direction), PU (parallel unknown), ST (stationary), and OT (others). To streamline this classification process, a sophisticated approach has been adopted.

Initially, a RoBERTa model, was employed to analyze the crash narratives. The primary objective was to categorize them into two overarching classes: crossing maneuvers (CR, CL, CU) and non-crossing maneuvers (PS, PO, PU, ST, OT). Any narrative that received a classification probability below 0.9 from the RoBERTa model was flagged for further scrutiny. In this subsequent stage, an SVM model was utilized, but this time, it was applied to the associated crash diagrams. This multi-step approach enables the prioritization of narratives that may require more detailed investigation based on the initial NLP classification.

The goal of SVM is to map the boxes' coordinates from the YOLOv5 output into a three-dimensional plane and find such a plane that can divide the coordinates representing the eight pedestrian maneuver classes.

SVM classifiers were then built using the "Caret" R package (56). Instead of splitting into training and test datasets, this study relied on 3 repeats of 10-fold cross-validation to estimate the test error. This study investigated SVM as classifiers for pedestrian maneuvers, vehicle maneuvers, and crash types, in different tuning parameters and in different kernel functions for best performances.

Results and Findings

First, a set of comparisons between linear and radial basis kernel functions were computed to see which SVMs performed better. SVMs with a radial basis kernel performed better than those SVMs with a linear kernel regardless of crossing or non-crossing classes. Moreover, SVMs were compared in different tuning parameter values for better SVM performance. Here, the tuned parameter is C, a regularization parameter which represents the trade-off between enlarging margin and lowering misclassification error. Because there is no rule of thumb in selecting a C value, thus different values were tested as shown in Table 19 below, with the C value and the highest accuracy results in bold.

Table 19. Accuracy Table for SVM Classifiers in Different Kernel Functions (Linear vs Radial).

PEDESTRIAN MANEUVER							
Crossing (N= 600)				Non-crossing (N = 400)			
Linear SVM		Radial SVM		Linear SVM		Radial SVM	
C	Accuracy	C	Accuracy	C	Accuracy	C	Accuracy
0.1052632	0.4332612	0.25	0.4521627	**0.1052632**	**0.3081288**	0.25	0.312048
0.2105263	0.4332612	0.5	0.4556015	0.2105263	0.3081288	0.5	0.3233689
0.3157895	0.4332612	1	0.4586793	0.3157895	0.3081288	1	0.3239131
0.4210526	0.4332612	2	0.4604072	0.4210526	0.3081288	2	0.3221592
0.5263158	0.4332612	4	0.4569679	0.5263158	0.3081288	4	0.3220133
0.6315789	0.4332612	8	0.4622109	0.6315789	0.3081288	8	0.3225542
0.7368421	0.4332612	16	0.4614895	0.7368421	0.3081288	16	0.3253872
0.8421053	0.4334429	32	0.4638419	0.8421053	0.3081288	32	0.3279399
0.9473684	0.4333521	64	0.466737	0.9473684	0.3081288	64	0.3330523
1.0526316	0.4334429	128	0.4665551	1.0526316	0.3081288	128	0.3422063
1.1578947	**0.4335337**	256	0.4708011	1.1578947	0.3081288	256	0.3524519
1.2631579	0.4332612	512	0.4776702	1.2631579	0.3081288	512	0.356912
1.3684211	0.4332612	1024	0.4830993	1.3684211	0.3081288	1024	0.3624522
1.4736842	0.4329888	2048	0.483279	1.4736842	0.3081288	2048	0.3659578
1.5789474	0.4328979	**4096**	**0.4844522**	1.5789474	0.3081288	**4096**	**0.369708**
1.6842105	0.4328979			1.6842105	0.3081288		
1.7894737	0.4328979			1.7894737	0.3081288		
1.8947368	0.4328979			1.8947368	0.3081288		
2	0.4328979			2	0.3081288		

Researchers then applied the best performing SVMs to the crossing and non-crossing crashes with probability values < 0.9, to see if they could be further subclassified by pedestrian maneuvers (CL, CR, CU and OT, PO, PS, PU, and ST, respectively). However, this resulted in low accuracy scores as shown in the first two data rows of Table 20. When researchers tested the SVMs on the crossing and non-crossing crashes with probability values > 0.9, the resulting accuracy scores were higher as shown in the next two data rows of Table 20. This suggests that further subclassification of pedestrian maneuvers (CL, CR, CU and OT, PO, PS, PU, and ST, respectively) may be possible with this approach by fine-tuning the parameters. The bottom two data rows of Table 20 provide the accuracy scores from the overall training dataset.

Data to describe pedestrian-vehicle crashes are available, but extracting the key details from the data is labor-intensive. These efforts should be supplemented with automated tools like computer vision and natural language processing to make large-scale analysis feasible.

Table 20. Accuracy of SVM Subclassification of Crossing and Non-Crossing Crashes.

P-VALUE	RADIAL SVM MODEL	SUBCLASSES	NUMBER OF RECORDS	C	ACCURACY
< 0.9	Crossing	CL, CR, CU	382	128	0.498
	Non-crossing	OT, PO, PS, PU, ST	291	4096	0.370
≥ 0.9	Crossing	CL, CR, CU	218	512	0.589
	Non-crossing	OT, PO, PS, PU, ST	109	1024	0.569
All	Crossing	CL, CR, CU	600	4096	0.484
	Non-crossing	OT, PO, PS, PU, ST	400	4096	0.370

Binary Intentionality Classification

For the binary intentionality classification task, researchers expanded the roster of NLP models to include some accessible high performing models. The tests included variants of models with different pretrained data and network architectures. These new models included Legal BERT (57), DeBERTaV3 (58), and ALBERT (59). However, the best performing model was the XLM RoBERTa Large model (60) which is a variant of the RoBERTa model that has been pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. This model achieves substantial gains in accuracy and F1 scores on tasks such as cross-lingual natural language inference (XNLI), named entity recognition (NER), and question answering (QA).

Even though there were noticeable gains when using BCE loss function in multiclass classification (i.e., crash typing), the same cannot be said with the intentionality binary classification. The results indicate that the best approach was to first identify the best-performing model and then explore the use of BCE loss function to improve the model's overall performance but even at the cost of reducing the accuracy. This tradeoff is more justifiable as the number of classes is increased and/or the imbalance grows more extreme. Table 21 shows the models and variants tested and the key performance metric values.

The average F1 was the primary key performance indicator with accuracy as the secondary indicator. The small amount of imbalance in the intentionality data was not a point of concern in these series of tests and some early tests indicated that in the intentionality binary task the improvement in F1 value came at the cost of reducing the accuracy. Thus, researchers used the Categorical Cross-Entropy (CCE) in the tests. Using 1.00 E-05 learning rate that performed best in the previous series of tests, researchers limited the batch size to 32 due to the computational restrictions encountered in the larger models and to maintain consistency across tests. The results shown in Table 21 are based on 10% of the data that the model was not previously exposed to during the fine-tuning process. In the tests using RoBERTa Large XLM, the researchers observed the best performance with 70.9% accuracy and 62.7% Average F1 (bolded row in Table 21).

Table 21. Pedestrian Intentionality Models Tested.

MODEL	VARIANT	ACCURACY	AVERAGE F1	AVERAGE RECALL	AVERAGE PRECISION
ALBERT	Base	71.9%	46.5%	52.2%	73.4%
ALBERT	Large	70.9%	60.7%	60.1%	63.1%
ALBERT	XLarge	70.9%	41.5%	50.0%	35.4%
BERT	Base	69.3%	45.3%	50.5%	52.2%
BERT	Base Legal	70.9%	41.5%	50.0%	35.4%
BERT	Base	65.8%	56.3%	56.1%	56.9%
BERT	Large	70.9%	41.5%	50.0%	35.4%
DeBERTa	Base	73.9%	54.9%	56.7%	74.4%
RoBERTa	Base	70.9%	41.5%	50.0%	35.4%
RoBERTa	Base XLM	70.9%	41.5%	50.0%	35.4%
RoBERTa	Large	74.9%	62.3%	61.5%	70.8%
RoBERTa	Large XLM	70.9%	62.7%	62.2%	63.7%

Not all crashes coded as “pedestrian-vehicle” crashes involve people who intended to be pedestrians. The circumstances of these crashes will vary, so the proper countermeasures will also vary.

Because of the high accuracy and F1 measures, researchers applied the model to the entire dataset to estimate the number of intended vs. unintended pedestrian crashes. Table 22 shows that overall, 78% of crashes involved intended pedestrians and 22% involved unintended pedestrians (Figure 21). That suggests that up to a fifth of ‘pedestrian’ crashes may have involved a person who was at the scene but not intending to walk (i.e., unintended pedestrian). In fact, this percentage of unintended pedestrian crashes increases to 43.5% for access-controlled facilities (i.e., freeways) where pedestrians are not expected because they are legally prohibited. This finding is in line with a 2019 TxDOT Dallas study of pedestrian crashes which found that about half of fatal pedestrian crashes on City of Dallas freeways involved unintended pedestrians (61). This trend also seems to match some TxDOT district’s observations of pedestrian crashes during their monthly fatal review team meetings. Whereas unintended pedestrian crashes are only 20.3% on non-controlled access facilities which seems reasonable and is more in line with general expectations.

Table 22. Pedestrian Intentionality by Access Control Type.

ACCESS CONTROL	INTENDED	UNINTENDED	TOTAL
Controlled	183	141	324
Non-Controlled	3283	835	4118
Total	3466	976	4442
Percent of Total	78%	22%	100%

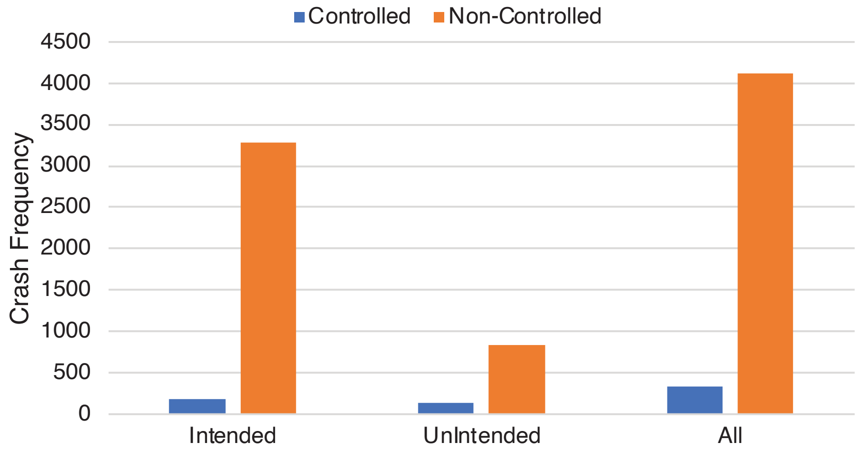


Figure 21. Pedestrian Intentionality by Access Control.

Weather

This task used the full dataset extracted from the CRIS database, which provides crash dates, times, direction of travel, fatality information, and other tidbits reported by the police. Sun glare and detailed weather data were not present in CRIS and needed to be captured from external sources before appending it to the crash data. This was a joint research effort by TTI and the NOAA Southern Regional Climate Center (SRCC). The sun glare and weather data analysis was conducted using SRCC data archives, which include a number of NOAA climate data sets.

Interpolation of Climate Elements

The first step involved estimating weather information at the location and time of the pedestrian crash. Automated Surface Observing Systems (ASOS) stations (62) were located in the three cities of the study as shown in Figure 22. These ASOS sites report weather elements such as temperature, precipitation, relative humidity/dewpoint, wind speed, cloud cover and visibility on an hourly basis. These data elements were captured using the Applied Climate Information Systems (ACIS) application programming interface (API) (63). The ASOS sites contributing to this project are

Overall, 22% of pedestrian-vehicle crashes involved unintended pedestrians-involved a person who was at the scene but not intending to walk. This estimate increases to 43.5% for access-controlled facilities. This finding is similar to a 2019 study which found that about half of fatal pedestrian crashes on City of Dallas freeways involved unintended pedestrians (61). This trend also seems to match some TxDOT district's observations of pedestrian crashes during their monthly fatal review team meetings.

listed in Table 23. Since the ASOS data occur at set points that may not be near the crash locations, the weather data must be interpolated across the entire metro area. An approach similar to another transportation related study (64) was used in this work to interpolate hourly climate information. For each of the numerical climate variables, a 0.1° x 0.1° interpolated grid (approximately 10,000 meters by 10,000 meters) was generated. Crash locations were mapped to grid points using the latitude-longitude data in CRIS. The climate element's value was captured and appended to the dataset. Because cloud cover values cannot be easily interpolated, the cloud cover reading reported at the nearest ASOS station was used.

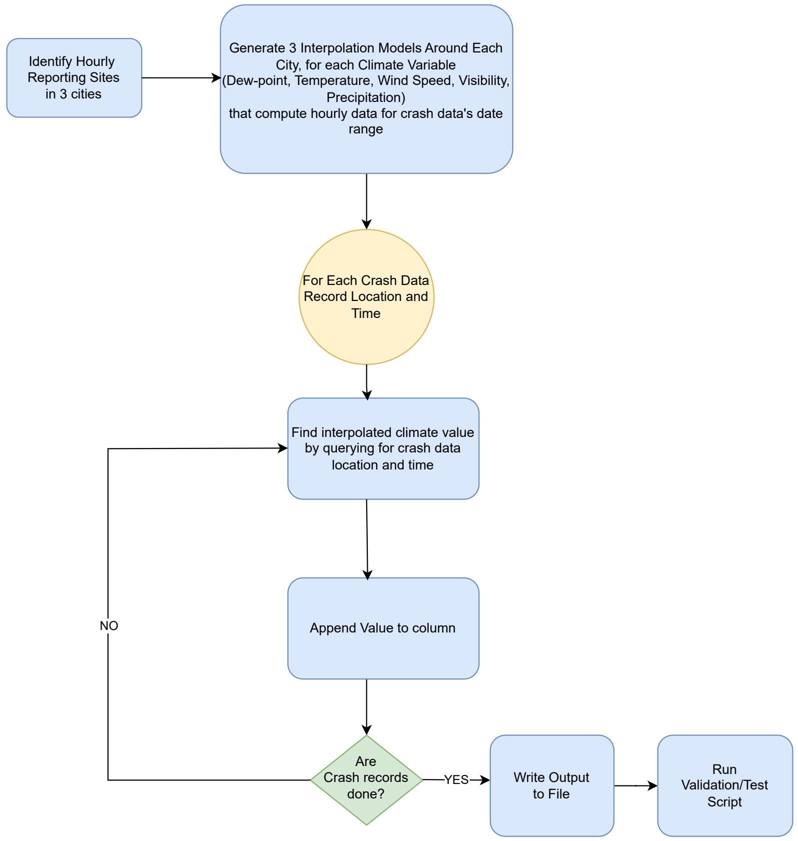

Identify Hourly Reporting Sites in 3 cities
Generate 3 Interpolation Models Around Each City, for each Climate Variable (Dew-point, Temperature, Wind Speed, Visibility, Precipitation) that compute hourly data for crash data's date range
For Each Crash Data Record Location and Time
Find interpolated climate value by querying for crash data location and time
Append Value to column
NO
Are Crash records done?
YES
Write Output to File
Run Validation/Test Script

Figure 22. Weather Data Interpolation Process.

In order to validate that the interpolated weather variables were feasible, researchers implemented a nearest-neighbor comparison with the known values at the ASOS stations. If interpolated temperature and dew point values differed by at least 5 F°, wind speeds differed by at least 5 mph, or visibilities differed by at least 2 mi, then we flagged the interpolation as suspicious. Similarly, interpolations that produced rainfall when the nearest ASOS station did not receive any (or vice versa) were also marked suspicious. As a result, 642 of the over 4000 entries received flags.

Table 23. ASOS Stations within the Study Area.

STATION ID	STATION NAME
KADS	Dallas Addison Airport
KATT	Austin-Camp Mabry
KAUS	Austin Bergstrom Intl AP
KDAL	Dallas FAA AP
KDFW	DAL-FTW WSCMO AP
KEDC	Austin Executive Airport
KF46	Rockwall Municipal Airport
KGPM	Grand Prairie Municipal Airport
KHQZ	Mesquite Metro Airport
KLNC	Lancaster Airport
KRBD	Dallas Redbird AP
KRND	Randolph AFB
KRYW	Lago Vista Rusty Allen Airport
KSAT	San Antonio Intl AP
KSKF	San Antonio Kelly Field AFB
KSSF	San Antonio Stinson Municipal AP

Climate API

The interpolated hourly gridded dataset was created for the same time span as the study period, 2018-2020. In order to query the dataset and to identify prevailing weather conditions near the crash location and time, an API was generated using similar principles as cited in work such as Jeffrey et al. (65). The API was developed using Python and the dataset was stored in the HDF5 format (66).

Determining the Presence of Sun Glare

The second step of the analysis required appending sun glare information at or near the reported crash time. Automated processes were developed based on NOAA's Solar Glare Calculator site and workbook (67). According to the site, NOAA states that the results of their calculations are theoretically accurate to within one minute for locations between 72°N and 72°S latitude. Since all the crashes lie within that band, these accuracies are acceptable. NOAA also discloses that official astronomical data are not kept, so these calculations are the best estimation available to researchers. Note that researchers decided to determine sun glare information based on data 10 minutes prior to the reported crash time. This temporal adjustment was made because the reported crash times are typically a little later than the time that the crash actually occurred, as per prior cited research (68). The researchers wanted to determine the glare conditions that were more likely present when the crash happened.

Before proceeding with the sun glare computation, it is necessary to define key angles. *Solar declination*, δ, is the angle between the equator and the sun at solar noon and can be thought of as the latitude at which the sun's rays are directly overhead. *Solar zenith angle*, θ_z, is the angle between the sun's rays and a vertical line drawn perpendicular to a site, such as a crash location. Solar zenith angle is the natural complement to solar elevation, α_s, which is the angle between the sun's rays and a plane tangential to the same site. Hour angle, ω, is the apparent azimuthal angle of the sun where the plane of reference is defined such that 0° represents solar noon, negative angles occur prior to solar noon, and positive angles occur after solar noon. Hour angle can be used to compute the true solar azimuth, γ_s. Direction of travel is provided in textual format in the CRIS data in 45° increments (north, southeast, etc.). Solar azimuth and direction of travel can be compared to determine whether "horizontal glare" may be possible at the time of crash.

Equation 8 provides the formula needed to compute the zenith angle, which can be translated to solar elevation using Equation 9. Equation 10 provides the formula needed to compute the solar azimuth. Based on the benchmarks established in prior research, glare may be present when $0° \leq \alpha_s \leq 45°$ and when γ_s is within 60° of the direction of travel (assumed from the unit record for vehicle 1 involved in the crash because pedestrian's direction of travel is often not available).

$$\theta_z = \cos^{-1}(\sin \emptyset \sin \delta + \cos \emptyset \cos \delta \cos \omega) \qquad (8)$$

$$\alpha_s = 90 - \theta_z \qquad (9)$$

$$\gamma_s = \begin{cases} \left[180 + \cos^{-1}\left(\dfrac{\sin \emptyset \cos \theta_z - \sin \delta}{\cos \emptyset \sin \theta_z}\right)\right] \bmod 360, & \text{if } \omega > 0 \\ \left[540 - \cos^{-1}\left(\dfrac{\sin \emptyset \cos \theta_z - \sin \delta}{\cos \emptyset \sin \theta_z}\right)\right] \bmod 360, & \text{otherwise} \end{cases} \qquad (10)$$

To demonstrate these calculations, researchers analyzed a crash that occurred in Austin at 12:39 PM local standard time on January 9, 2018. Solar elevation was calculated to be 37.7° above the horizon. Since elevations below 45° may result in glare, the researchers continue to the horizontal component. Solar azimuth was calculated to be 179° (relative to north). The crash happened when the driver was facing southwest, 225° from north. This means that the horizontal component was 46°, within the 60° threshold. Since neither the vertical nor horizontal components refute the presence, the researchers conclude that this crash could have been glare-related.

Results And Findings

Table 24 shows the distribution of crashes by the weather condition variable in CRIS and the cloud coverage variable provided by SRCC. The comparison reveals good agreement between the CRIS data and the SRCC data.

Figure 23 shows the distribution of crashes by the hourly precipitation variable from the SRCC hourly interpolated data. About 95.5% of the crashes occurred in hours with no precipitation, and the rest of the crashes occurred in hours with as much as 1 inch of precipitation. The data are plotted on a logarithmic y-axis to improve the visibility of the small percentages for the non-zero data points.

Table 24. Crash Distribution by Weather Condition and Cloud Coverage.

WEATHER CONDITION (CRIS)	PERCENT OF CRASHES BY CLOUD COVERAGE (SRCC)					
	Clear	Scattered	Broken	Overcast	Obscured	Total
Clear	38.20	29.84	6.97	3.94	0.05	79.00
Cloudy	2.39	5.10	2.81	3.55	0.03	13.88
Rain	0.13	2.00	1.63	2.50	0.05	6.31
Unknown	0.05	0.16	0.11	0.05	0.00	0.37
Fog	0.03	0.03	0.00	0.11	0.16	0.32
Sleet/Hail	0.03	0.00	0.00	0.03	0.00	0.05
Other	0.03	0.00	0.00	0.03	0.00	0.05
Blowing sand/snow	0.03	0.00	0.00	0.00	0.00	0.03
Total	40.88	37.12	11.51	10.20	0.29	100.00

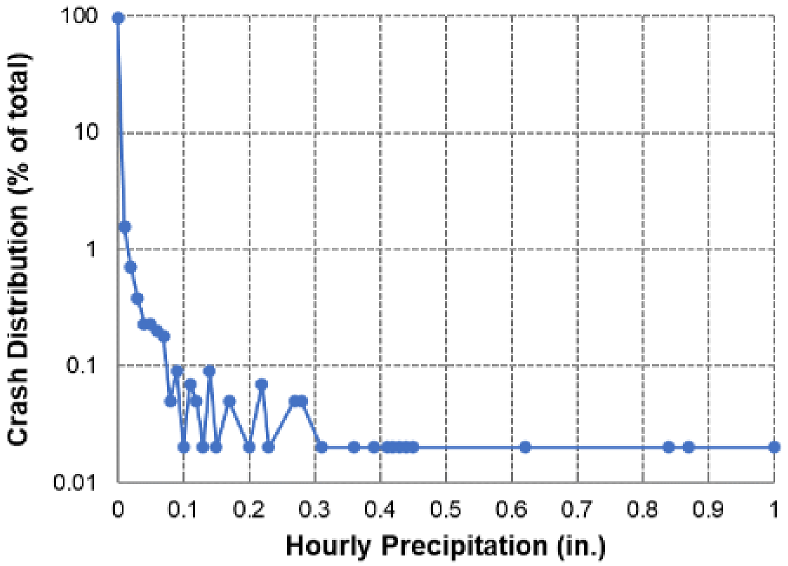


Figure 23. Distribution of Crashes by Precipitation (Logarithmic).

The CRIS crash data file contains a variable called “Othr_Factr_ID” which describes “other” factors related to crashes. This variable has 54 possible code values, including a value of 21 for “vision obstructed by headlight or sun glare”. There is only one “other factor” variable in the CRIS database, and it is tied to the crash, not units (i.e., vehicle or pedestrian). This variable can describe up to one additional factor that was not captured in the unit records but is deemed important by the reporting officer. While this information may provide insight into glare issues, its entry is not consistently applied.

Figure 24 shows the distribution of crashes by the “other factor” variable. A total of 25 crashes, or 0.56% of the sample, were coded as being affected by glare. The figure includes the five most common values for the crashes and the glare value, with the remaining values collapsed into an “other” category. As shown, the variable is not commonly used by officers – about two-thirds of the values are “not applicable”. The code values for entering or leaving driveways were among the common values used.

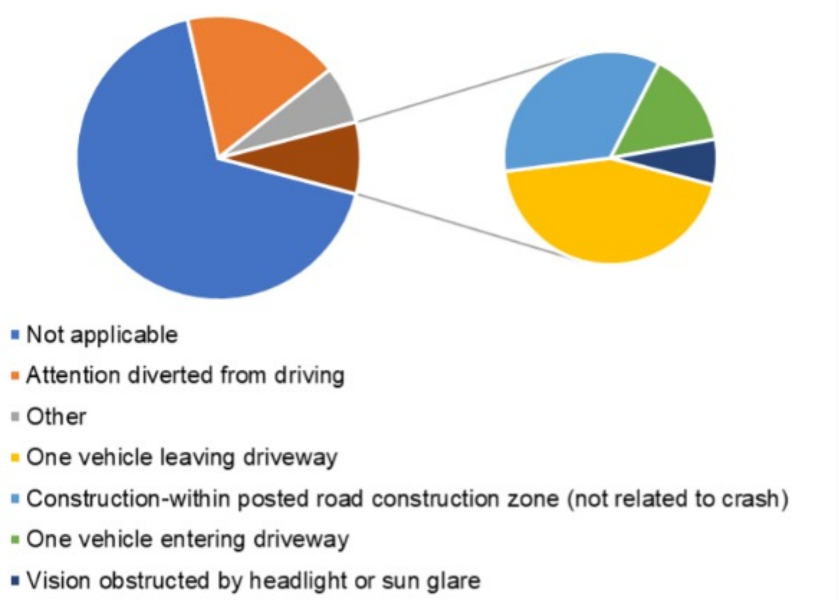


Figure 24. Distribution of Crashes by Other Factor Variable in CRIS.

Table 25 shows the distribution of crashes by the “other factor” variable from CRIS and the glare flag variable (binary 1/0 to indicate if sun glare was a possible factor in the crash) from the SRCC data. Note that when the counts of glare-related crashes and non-glare-related crashes (the second and third columns in the table, respectively) are compared, the non-glare-related crashes are more abundant in every row of the table except the row for crashes flagged for glare in the CRIS database. These counts are shown in bold in the table.

Table 25. Crash Distribution by Other Factor Variable and Sun Glare Flag.

OTHER FACTOR (CRIS)	SRCC SUN GLARE FLAG (BINARY 1/0)		
	Yes	No	Total
Not applicable	363	2568	2931
Attention diverted from driving	133	639	772
Other	41	250	291
One vehicle leaving driveway	28	126	154
Construction-within posted road construction zone (not related to crash)	19	103	122
One vehicle entering driveway	12	37	49
Vision obstructed by headlight or sun glare	**20**	**3**	**23**
Total	616	3726	4342

The hours with the highest probability of glare-flagged crashes occur near sunrise and sunset, which is when the sun sits lowest in the sky.

Figure 25 shows the distribution of glare-flagged crashes by hour. The distribution is consistent with the expected sun position and angle at the various hours of the day. Figure 26 shows the distribution of crashes by hour and glare flag. The hours with the highest probability of glare-flagged crashes occur near sunrise and sunset, which is when the sun sits lowest in the sky. Additionally, Figure 26 shows dips in non-glare crashes compared to other nearby hours, justifying the claim that glare is possibly a factor. These two figures provide perspective on the relative contribution of sun glare to the crash totals in the hours of the day.

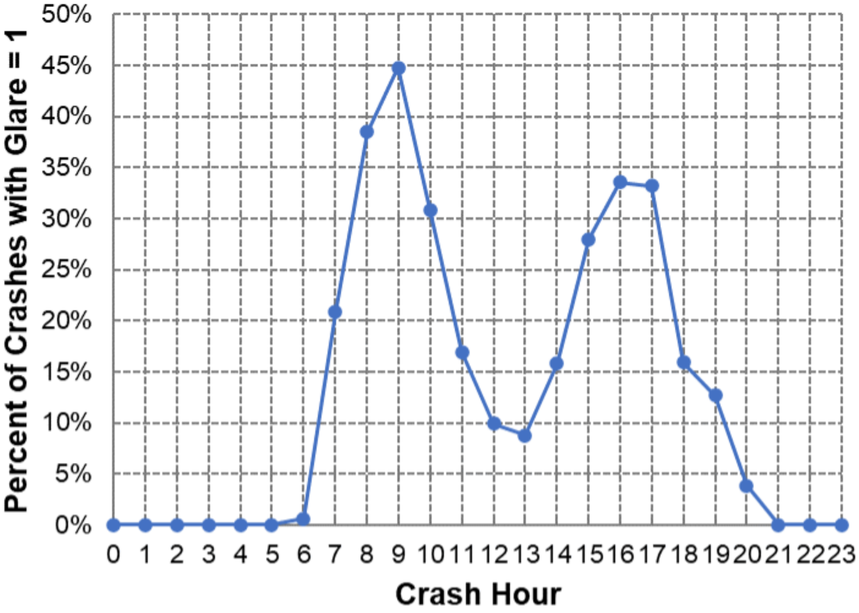


Figure 25. Distribution of Glare-Flagged Crashes by Hour.

The research team used a text querying tool to review the crash narratives for additional evidence of glare as a contributing factor to the crashes in the database. This effort included the following keywords: glare, sun, eye, bright, blind, shadow, shine, and shining. The text querying tool produced a list of crash ID numbers for each keyword, and the research team read the narratives for these crashes to determine if the crash was related to sun glare (i.e., to ensure that the keyword was referring to sun glare instead of another issue or an unrelated observation). Table 26 provides a count of crashes that had the various keywords in their narratives that referred to sun glare and/or a code value of 21 ("vision obstructed by headlight or sun glare") in the "other factor" variable, compared with the value of the binary sun glare flag from the SRCC data. The keywords "sun" and "eye" were the two most common keywords.

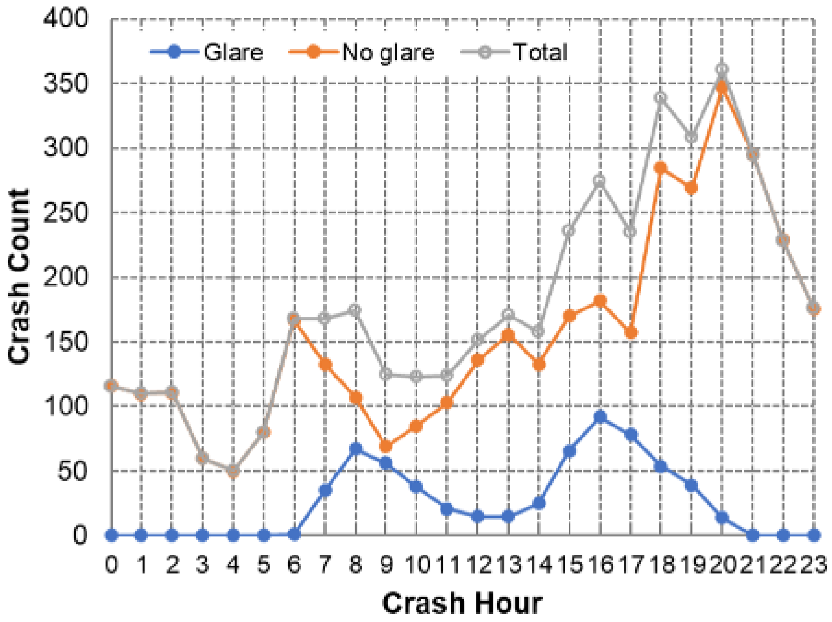


Figure 26. Distribution of Crashes by Hour and Glare Flag.

Table 26. Comparison of Keywords, Factor Codes, and Sun Glare Flag.

KEYWORD OR CODE VALUE IN CRIS	CRASH COUNT		
	All Crashes	Sun Glare Flag = 1	Sun Glare Flag = 0
Glare	8	5	3
Sun	58	46	12
Eye	30	27	3
Bright	5	3	2
Blind	11	9	2
Shadow	1	1	0
Shine	1	1	0
Shining	3	3	0
"Other factor" variable = 21	25	20	5
Total	**142**	**115**	**27**

Some of the identified crashes had more than one of the keywords in their narratives and/or were also flagged as glare-related with the "other factor" variable. Table 27 provides a count of crashes by number of CRIS-related flags and the sun glare flag from the SRCC analysis. A total of 65 crashes had some evidence of sun glare based on the CRIS data, and a total of 616 crashes were identified as possibly glare-related in the SRCC data. A total of 630 crashes (14.2% of the total) had evidence of sun glare in the CRIS data or were found to be possibly glare-related in the SRCC analysis.

As a result, 630 crashes (14.2%) were found to be possibly glare-related.

Table 27. Count of Crashes by Glare Evidence in CRIS and Appended Data.

NUMBER OF KEYWORDS OR FLAGS IN CRIS	CRASH COUNT		
	All Crashes	Sun Glare Flag = 1	Sun Glare Flag = 0
0	4377	565	3812
1	15	10	5
2	29	24	5
3	16	12	4
4	4	4	0
5	1	1	0
Total	**4442**	**616**	**3826**

A total of 565 crashes (or 12.7% of the total) were flagged as possibly glare-related in the SRCC data analysis but did not have any mention of sun glare in their narratives or in the “other factor” variable from CRIS. There are two explanations for this discrepancy. Some of these crashes may not have been affected by sun glare even though they occurred in conditions (time of day, sun position, direction of travel, etc.) when sun glare could have been a contributing factor. Additionally, some of these crashes may have been glare-related but the glare issue did not get identified or recorded in the crash investigation and documentation process.

Crash Severity Predictive Analysis

Using the manually reviewed training dataset, researchers developed a regression model to see what factors could influence the severity of pedestrian crashes, particularly for intended pedestrians on controlled access facilities versus pedestrians on non-controlled access facilities. Note, this analysis was not performed on the full dataset because PBCAT crash types could not be estimated accurately.

Table 28 shows that there were 66 intended pedestrian crashes on controlled-access facilities and 637 on non-controlled access. Researchers conflated these crashes to TxDOT's RHINO network to capture traffic and geometric factors such as maximum speed (posted regulatory speed limits), geometry, ADTs, number of lanes, etc. Researchers also considered the PBCAT crash type and weather information such as visibility, temperature, precipitation, sun glare, etc.

Table 28. Distribution of Crashes by Severity and Access Control.

SEVERITY	CONTROLLED ACCESS		NON-CONTROLLED ACCESS	
	Frequency	Percent	Frequency	Percent
K	46	69.7	128	20.13
A	14	21.21	189	29.72
B	2	3.03	173	27.2
C	1	1.52	130	20.44
O	3	4.55	16	2.52
Total	**66**	**100**	**637**	**100**

Researchers calibrated logistic regression models to estimate the probability of fatality for pedestrian-vehicle crashes on controlled-access roadways and fatal and incapacitating injury crashes on non-controlled-access roadways. A logistic regression model expresses the log-odds of the probability of a crash occurrence as a function of speed and other roadway characteristic variables. The model is described as follows:

$$g(\boldsymbol{x}) = ln\left[\frac{P(Y_{it} = 1|\boldsymbol{x})}{1 - P(Y_{it} = 1|\boldsymbol{x})}\right] = \beta_0 + \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it} \tag{11}$$

$$P(Y_{it} = 1|\boldsymbol{x}) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{\beta_0 + \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it}}}{1 + e^{\beta_0 + \beta_1 x_{1,it} + \cdots + \beta_K x_{K,it}}} \tag{12}$$

where:

$P(Y_{it})$ = probability of high-severity crash occurrence on segment *i* in time *t***,** given a crash has occurred.

$g(\boldsymbol{x})$ = logit (log-odds).

$\boldsymbol{x}$ = vector of predictor variables.

x_i = predictor variable representing a site condition.

β_i = regression coefficient.

Note that the underlying assumption for the above model is that the relationship between the logit, *g(x)*, and predictor variables is linear. The intercept β_0 represents the baseline level of the logit, and β_k represents the change in the logit that occurs with a unit change in X_k.

Table 29 provides the model calibration results for estimating the probability of fatal pedestrian-vehicle crash on a controlled-access roadway. Note the small sample size did not allow for predicting all crash severities. Positive coefficient values show that the probability of fatality increases with higher truck AADT, if the maneuver type is S_CL (motorist going straight, pedestrian crossing from motorist's left), or if the pedestrian is male. The latter two coefficients likely represent drivers' attention focus (i.e., they are less likely to expect a pedestrian crossing from their left than from their right) and different risk-taking behavior for men compared to women.

Table 29. Maximum Likelihood Estimates for High-Severity Crash Probability on Controlled-Access Roadway.

VARIABLE	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PROBABILITY > WALD CHI-SQUARE
Intercept	-1.5026	0.8800	2.9155	0.0877
Truck AADT	0.000092	0.000052	3.1833	0.0744
Maneuver type S_CL	1.5613	0.8641	3.2646	0.0708
Pedestrian gender male	1.9756	0.6401	2.8233	0.0929

Table 30 provides the model calibration results for estimating the probability of fatality or incapacitating injury (K or A) for a pedestrian-vehicle crash on a non-controlled-access roadway. Positive coefficient values show that the probability of K or A fatality increases with increasing outside shoulder width, visibility, or pedestrian age, or if the maneuver type is S_CL or S_CR (motorist going straight, pedestrian crossing from motorist's left or right), and the probability decreases if sun glare is present. The results for visibility and sun glare presence likely correlate with vehicle speeds; that is, drivers may choose higher speeds with greater visibility and absence of sun glare. The coefficients for both S_CL and S_CR maneuver types are positive, but the coefficient value for S_CL is greater, again likely reflecting drivers' tendency to expect crossing pedestrians to originate from their right (i.e., the roadside or the near-side sidewalk) more often than from their left. The positive coefficient value for outside shoulder width may reflect other roadway characteristics or correlate with vehicle speeds but does not suggest that a wider outside shoulder in itself will increase the severity distribution of pedestrian-vehicle crashes.

Fatality increases with higher truck traffic, if the maneuver type is S_CL (motorist going straight, pedestrian crossing from motorist's left), or if the pedestrian is male, likely due to drivers' attention focus and different risk-taking behavior for men compared to women.

Fatality or serious injury increases with visibility and decreases if sun glare is present, likely because these conditions affect vehicle speeds. Other relevant variables include outside shoulder width (which correlates with other design variables), pedestrian age, and maneuver type (especially S_CL and S_CR).

Table 30. Maximum Likelihood Estimates for High-Severity Crash Probability on Non-Controlled-Access Roadway.

VARIABLE	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PROBABILITY > WALD CHI-SQUARE
Intercept	-2.7803	0.5680	23.9587	< 0.0001
Outside shoulder width	0.0570	0.0179	10.1716	0.0014
Maneuver type S_CL	1.1076	0.2327	22.6477	< 0.0001
Maneuver type S_CR	0.7369	0.2328	10.0243	0.0015
Visibility	0.1673	0.0541	9.5617	0.0020
Sun glare presence	-0.5595	0.2815	3.9502	0.0469
Pedestrian age	0.0152	0.00483	9.9352	0.0016

The calibrated crash severity models are described as follows:

$$P_{K|c} = \frac{e^{-1.5026+0.000092v_{tk}+1.5613I_{S\_CL}+1.0756I_M}}{1+e^{-1.5026+0.000092v_{tk}+1.5613I_{S\_CL}+1.0756I_M}} \qquad (13)$$

$$P_{KA|nc} = \frac{e^{-2.7803+0.057w_s+1.1076I_{S\_CL}+0.7369I_{S\_CR}+0.1673d_{vis}-0.5595I_g+0.0152A_p}}{1+e^{-2.7803+0.057w_s+1.1076I_{S\_CL}+0.7369I_{S\_CR}+0.1673d_{vis}-0.5595I_g+0.0152A_p}} \qquad (14)$$

where:

$P_{K|c}$ = probability of fatality (K) in a pedestrian-vehicle crash on a controlled-access roadway, given a crash has occurred.

$P_{KA|nc}$ = probability of fatality or incapacitating injury (KA) in a pedestrian-vehicle crash on a non-controlled-access roadway, given a crash has occurred.

v_{tk} = truck AADT (trucks/day).

I_{S_CL} = indicator variable for S_CL maneuver type (=1 if the maneuver is classified as S_CL, 0 otherwise).

I_{S_CR} = indicator variable for S_CR maneuver type (=1 if the maneuver is classified as S_CR, 0 otherwise).

I_M = indicator variable for male gender (=1 if the pedestrian is male, 0 otherwise).

w_S = outside shoulder width (ft).

d_{vis} = visibility distance (mi).

I_g = indicator variable for sun glare presence (=1 if present, 0 otherwise).

A_p = pedestrian age (years).

CONCLUSIONS

This study sought to explore and apply advanced techniques in NLP, CV, and machine learning to data mine the unstructured data in crash reports for vulnerable road users, particularly pedestrians. More specifically, researchers sought ways to extract pedestrian maneuvers and intentionality from the crash narratives and diagrams. Alternate data sources were also mined for weather information to 'fill the gap' of key climate factors as well as sun glare.

With regards to multiclassification, this study addressed data imbalance issues commonly encountered in transportation data classification. The researchers tested two classification systems based on PBCAT 3.0. Increasing the dataset size or choosing a smaller classification system with more distinguishable sentiments improved the model's performance. Further improvements were achieved by using the Balanced Categorical Cross Entropy loss function and a more robustly pretrained model. Researchers first tested the capability of models using eight classes with the highest accuracy of 34% and highest average F1 of 25%. These tests did not demonstrate a significant advantage in using BCE loss as a strategy to improve the model performance. However, researchers then performed tests with a more abstract classification system comprised of four classes. In RoBERTa Base tests on the abstract classification system with four classes, researchers observed the best performance was 65.2% accuracy and 50.9% Average F1. The improvement in F1 performance was consistently observed in models using the BCE loss function. Thus, multiclassification with smaller number of classes can yield more accurate results and the BCE loss function can yield a more balanced result. Researchers then decided to test combining NLP with CV and machine learning. When researchers tested the SVMs on a two-class system: crossing and non-crossing crashes with probability values > 0.9, the resulting accuracy scores improved. This suggests that further subclassification of pedestrian maneuvers (CL, CR, CU and OT, PO, PS, PU, and ST, respectively) may be possible with this approach by fine tuning the parameters.

Pedestrian safety treatments' effectiveness will vary significantly depending on the circumstances that lead people to become pedestrians (i.e., walking to a destination versus exiting a stalled vehicle), so treatment decisions should be based on the most accurate information as possible.

Researchers tested the binary intentionality classification using multiple LLMs. They observed the best performance with the In the RoBERTa Large XLM tests, with 70.9% accuracy and 62.7% Average F1. Because of the high accuracy and F1 measures, researchers applied the model to the entire dataset to estimate the number of intended vs. unintended pedestrian crashes. Overall, 78% of crashes involved intended pedestrians and 22% involved unintended pedestrians. That suggests that up to a fifth of 'pedestrian' crashes may have involved a person that was at the scene but not there intending to walk (i.e., unintended pedestrian). In fact, this

TX Transportation Code S 5551.351(2) defines a pedestrian as any person who is not an occupant of a motor vehicle in transport. This definition fails to distinguish between intended pedestrians (as defined in this study), who may benefit from safety treatments, and unintended pedestrians, who likely will not.

percentage of un increases to 43.5% for access-controlled facilities (i.e., freeways) where one would not expect to encounter pedestrians because they're legally prohibited to be on them. This finding is in line with a 2019 TxDOT Dallas study of pedestrian crashes which found that about half of fatal pedestrian crashes on City of Dallas freeways were unintended. This trend also seems to match some TxDOT district's observations of pedestrian crashes during their monthly fatal review team meetings. Whereas unintended pedestrian crashes are only 20.3% on non-controlled access facilities which seems reasonable and is more in line with general expectations. This finding suggests that if a pedestrian's intentionality is considered, then 'pedestrian' crashes may be overcounted in crash statistics. However, part of the problem may be how pedestrians are defined. According to TX Transportation Code: S 5551.351(2), a pedestrian is any person who is not an occupant of a motor vehicle in transport. This legal definition includes motorized and non-motorized wheelchairs. But this definition also includes persons that are struck while outside of vehicle, off a bike/ motorcycle, on e-scooters or construction workers, emergency responders, etc., who are struck while at the scene even though their 'intention' is not to walk along or across the roadway. This issue deserves more consideration as transportation agencies continue to rely on these crash statistics to analyze, plan, develop, and ultimately allocate resources to addressing pedestrian safety. Pedestrian safety treatments' effectiveness will vary significantly depending on the circumstances that lead people to become pedestrians so decisions should be based on the most accurate information as possible.

Weather information, including sun glare, is often limited or not available in crash reports. It is dependent on the officer noting the data in the crash report. Moreover, it may be inaccurate as the officer may report weather conditions observed at a time long after the crash, or not report it at all. Other data sources are sometimes needed to fill this data gap to understand if weather was a potential contributing factor to the crash. This study developed a technique to address this gap in weather and solar glare information. The technique interpolates available climate data to gain a better estimate of the weather conditions at a specific time and place (coordinate), and the presence of sun glare. The work also developed a climate data API for the hourly interpolated information that can be used for future research work and to append to other crash data sets. Some key findings of this research effort included:

- The weather condition variable in CRIS and the cloud coverage variable provided by SRCC indicates a good agreement between the CRIS data and the SRCC data analysis.

- About 95.5% of the crashes occurred in hours with no precipitation, and the rest of the crashes occurred in hours with as much as 1 inch of precipitation. This was close to the percentage of crashes flagged for "rain" which was 6.3%.
- The results show that 630 (14.2%) of the pedestrian crashes could have been affected by sun glare. Sixty-five of them had some evidence of sun glare based on the CRIS data whereas 616 were possibly glare-related based on the SRCC data. The distribution of glare-related crashes by hour is consistent with the expected sun position and angle throughout the day.
- In this analysis, 565 of the crashes (or 12.7% of the total) were flagged as possibly glare-related in the SRCC data but did not have any mention of sun glare in their narratives or in the "other factor" variable from CRIS. There is a potential for glare being an under-reported contributing factor in the crash reports. Some of these crashes may not have been affected by sun glare even though they occurred in conditions (time of day, sun position, direction of travel, etc.) when sun glare could have been a contributing factor. Additionally, some of these crashes may have been glare-related but the glare issue did not get identified or recorded in the crash investigation and documentation process.

The weather condition variable in CRIS and the cloud coverage variable provided by SRCC indicates a good agreement between the CRIS data and the SRCC data analysis.

Application to Texas Transportation Safety

Overall, this effort highlights how important it is to utilize advanced techniques to extract key, unstructured information from pedestrian-vehicle crash data. Manual investigations are a labor-intensive process, so practitioners should consider using tools like natural language processing and computer vision to understand what is happening and be able to identify appropriate treatments, if any. The tools and techniques explored in this study can be used to broaden our knowledge of traffic crash details, focusing on pedestrian-vehicle crashes where details are often limited. Reviewing and interpreting crash report narratives and diagrams manually would be time-prohibitive on a large scale such as all of Texas. Moreover, researchers often calibrate cross-sectional models to predict crash frequency, but these models are limited because it is expensive to build a database to describe the crash circumstances in detail.

The key questions that this research effort helped to answer are:

1. Was the involved pedestrian actually intending to be a pedestrian?
2. If so, what caused the crash from the perspective of both units?
3. What countermeasures are justified to address the causes in the future?

This type of information can be used to support policy or design decisions by practitioners such as at TxDOT, municipalities, or pedestrian safety coalitions.

FUTURE RESEARCH

Although the contributions from this study are impactful, results show that there are areas that could benefit from further research. Some examples are described below.

Although the contributions from this study are impactful, results show that there are areas that could benefit from further research.

Natural Language Processing

This research has revealed some challenges in developing more complex crash typing models. The limited size of the datasets, inconsistent narration styles, and extreme data imbalance are identified as significant factors hindering progress. To address these issues, it is recommended to increase the size of training datasets, particularly for specific crash types to achieve better balance; to utilize other large language models, such as ChatGPT, with greater exposure to crash narratives or similar reports during their pre-training stage. This exposure will enhance the models' understanding of the nuances and complexities present in the crash narratives, leading to more accurate and reliable results. Finally, adopting cross-validation evaluation methods in future studies to mitigate inconsistencies observed in evaluating the multiclassification task. Cross-validation helps ensure the model's performance is thoroughly assessed and not overly influenced by random variations in the data-splitting process.

Computer Vision & Machine Learning

This research shows there is potential in applying CV and machine learning to crash diagrams as part of a multistep, multiclassification approach. For example, the size of testing dataset should be increased, particularly for specific crash types to achieve better balance, and/or limiting the crash diagrams to only those that contain at least the four classes required. Exploring new augmentation methods can also help to increase dataset size. Finally, exploring other supervised learning models with associated algorithms for classification is also recommended.

Predictive Analysis

The predictive analysis was limited by the small sample size once the training dataset was filtered for intended pedestrians. Ideally, this analysis should be repeated utilizing the full dataset once the PBCAT crash types can be predicted with an acceptable level of accuracy. Secondly, it would help if the crashes could be conflated with other datasets such as the HERE (formerly Navteq) network for more accurate regulatory speed limits, intersection control such as traffic signals or stop-control, transit amenities such as bus and light rail stops, presence of sidewalks, crosswalks, and roadway lighting, etc.

Having specific weather data such as rainfall intensity and risk of solar glare can provide additional insights on how weather and/or solar glare may have played a role in the crash.

Weather

The research showed that while some weather information in crash reports can be corroborated with interpolated weather data, having specific weather data such as rainfall intensity and risk of solar glare can provide additional insights on how weather and/or solar glare may have played a role in the crash, particularly for pedestrian related crashes. However, it also identified two specific areas for further research. Firstly, more work could be done in refining the interpolation and analysis to incorporate sub-hourly climate data. This would further define the conditions at the specific time and location in question. New data being made available will also better consider the variability of climate conditions between weather recording stations. Secondly, future work can involve improving the accuracy of the interpolated climate information. Quality control on the interpolated climate information yielded 3719 higher confidence records and 642 lower confidence records for the 4361 crashes with sufficient weather data. This can be improved by incorporating additional information such as radar-informed precipitation information.

ACKNOWLEDGEMENTS

Researchers would like to acknowledge and thank the Texas A&M Transportation Institute's Center for Transportation Safety for funding this effort. The TTI team is grateful to the data scientists and engineers at NOAA Southern Regional Climate Center and Trabus Technologies for providing the cloud and data infrastructure to conduct this study. Lastly, portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

REFERENCES

1. NHTSA. Traffic Safety Facts - 2019 Data: Pedestrians. U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, D.C., 2021.

2. Beck, L. F., A. M. Dellinger, and M. E. O'Neil. Motor Vehicle Crash Injury Rates by Mode of Travel, United States: Using Exposure-Based Methods to Quantify Differences. In American Journal of Epidemiology, Vol. 166, 2007, pp. 212–218.

3. NHTSA. Fatality Analysis Reporting System (FARS). https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. Accessed October 12, 2023.

4. Thomas, L., D. Levitt, M. Vann, K. Blank, K. Nordback, and A. West. PBCAT–Pedestrian and Bicycle Crash Analysis Tool Version 3.0. Federal Highway Administration, Washington, D.C., 2021.

5. Federal Highway Administration. Develop Pedestrian and Bicycle Crash Analysis Tool. In Public Roads, Vol. 63, No. 5, 2000, p. 54.

6. Shah, N. R., S. Aryal, Y. Wen, and C. R. Cherry. "Comparison of Motor Vehicle-Involved e-Scooter and Bicycle Crashes Using Standardized Crash Typology". In Journal of Safety Research, Vol. 77, 2021, pp. 217–228. https://doi.org/10.1016/j.jsr.2021.03.005.

7. Lopez, D., L. C. Malloy, and K. Arcoleo. "Police Narrative Reports: Do They Provide End-Users with the Data They Need to Help Prevent Bicycle Crashes?" In Accident Analysis & Prevention, Vol. 164, 2022. https://doi.org/10.1016/j.aap.2021.106475.

8. Schneider, R. J., and J. Stefanich. "Application of the Location–Movement Classification Method for Pedestrian and Bicycle Crash Typing". In Transportation Research Record: Journal of the Transportation Research Board, No. 2601, 2016, p. pp 72-83.

9. Chavis, C., Y.-J. Lee, and S. Dadvar. Analysis of Bicycle and Pedestrian Crash Causes and Interventions. 2018, p. 368p.

10. NHTSA. 2020 FARS/CRSS Pedestrian Bicyclist Crash Typing Manual: A Guide for Coders Using the FARS/CRSS Ped/Bike Typing Tool. Washington, D.C., 2022, p. 86p.

11. Banks, G. C., H. M. Woznyj, R. S. Wesslen, and R. L. Ross. "A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)". In Journal of Business and Psychology, Vol. 33, No. 4, 2018, pp. 445–459.

12. Kwayu, K. M., V. Kwigizile, J. Zhang, and J.-S. Oh. "Semantic N-Gram Feature Analysis and Machine Learning–Based Classification of Drivers' Hazardous Actions at Signal-Controlled Intersections". In Journal of Computing in Civil Engineering, Vol. 34, No. 4, 2020, p. 04020015.

13. Das, S., M. Le, and B. Dai. "Application of Machine Learning Tools in Classifying Pedestrian Crash Types: A Case Study". In Transportation Safety and Environment, Vol. 2, No. 2, 2020, pp 106-119. https://doi.org/10.1093/tse/tdaa010.

14. Oliaee, A., S Das, J Liu, MA Rahman. "Using Bidirectional Encoder Representations from Transformers (BERT) to Classify Traffic Crash Severity Types". In Natural Language Processing Journal 3, 100007, 2023.

15. Weng, Y., S Das, SG Paal. "Applying Few-Shot Learning in Classifying Pedestrian Crash Typing". In Transportation Research Record, 03611981231157393, 2023.

16. Das, S., A. Oliaee, M. Le, M. Pratt, and J Wu. Classifying Pedestrian Maneuver Types using the Advanced Language Model. In Transportation Research Record, 03611981231155187, 2023.

17. Kwayu, K. M., V. Kwigizile, K. Lee, and J.-S. Oh. "Discovering Latent Themes in Traffic Fatal Crash Narratives Using Text Mining Analytics and Network Topology". In Accident Analysis & Prevention, Vol. 150, 2021, p. 105899.

18. Ultralytics/yolov5: v3.0. https://zenodo.org/record/3983579. Accessed October 2, 2023.

19. Wang, C., A. Bochkovskiy, and H. Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. https://arxiv.org/abs/2207.02696. Accessed October 2, 2023.

20. Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

21. He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition. In “Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition”, 2016, pp. 770-778.

22. He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask r-Cnn. In “Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition”, 2017, pp. 2961-2969.

23. Zhou, Z., Li, Y., Li, J., Yu, K., Kou, G., Wang, M., & Gupta, B. B. Gan-Siamese Network for Cross-Domain Vehicle Re-Identification in Intelligent Transport Systems. IEEE Transactions on Network Science and Engineering, 2022.

24. Narayanan, A., Kumar, R. D., Roselin Kiruba, R., & Sharmila, T. S. Study and Analysis of Pedestrian Detection in Thermal Images using YOLO and SVM. In Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), IEEE, March 2021, pp. 431-434.

25. Kolla, E., Adamová, V., & Vertal', P. Simulation-Based Reconstruction of Traffic Incidents from Moving Vehicle Mono-Camera. In Science & Justice, Vol. 62, No. 1, 2022, pp. 94-109.

26. Deng, Z., Chen, Y., Liu, L., Wang, S., Ke, R., Schonlieb, C. B., & Aviles-Rivero, A. I. TrafficCAM: A Versatile Dataset for Traffic Flow Segmentation. arXiv preprint arXiv:2211.09620, 2022.

27. Doshi, K., & Yilmaz, Y. An Efficient Approach for Anomaly Detection in Traffic Videos. In “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, 2021, pp. 4236-4244.

28. Yuan, Y., Wang, D., & Wang, Q. Anomaly Detection in Traffic Scenes via Spatial-Aware Motion Reconstruction. In IEEE Transactions on Intelligent Transportation Systems, Vol. 18, No. 5, 2016, pp. 1198-1209.

29. Ma, H., P. Chen, S. Chen, L. Chen, V. Linkov, and C. Pai. Population-Based Case-Control Study of the Effect of Sun Glare on Pedestrian Fatalities in Taiwan. In BMJ Open, Vol. 9, 2019.

30. Mitra, S. Sun Glare and Road Safety: An Empirical Investigation of Intersection Crashes. In Safety Science, Vol. 70, 2014, pp. 246-254.

31. Hagita, K. and K. Mori. “The Effect of Sun Glare on Traffic Accidents in Chiba Prefecture, Japan. In Asian Transport Studies, Vol. 3, No. 2, 2014, pp. 205–219.

32. Billah, K., H. O. Sharif, and S. Dessouky. “Analysis of Pedestrian–Motor Vehicle Crashes in San Antonio, Texas”. In Sustainability, Vol. 13, No. 12, 2021, p. 6610.

33. Li, X., B. Yang, W. Qiu, J. Zhao, and C. Ratti. A Novel Method for Predicting and Mapping the Occurrence of Sun Glare using Google Street View. In Transportation Research Part C, Vol. 106, 2019, pp. 132-144.

34. Zhao, B., N. Zuniga-Garcia, L. Xing, and K. M. Kockelman. “Predicting Pedestrian Crash Occurrence and Injury Severity in Texas using Tree-Based Machine Learning Models. In Transportation Planning and Technology, June 2023, pp. 1–22.

35. Charm, T., H. Wang, N. Zuniga-Garcia, M. Ahmed, and K. M. Kockelman. Predicting Crash Occurrence at Intersections in Texas: An Opportunity for Machine Learning. In Transportation Planning and Technology, February 2023, pp. 1–22.

36. Das, S., R. Tamakloe, H. Zubaidi, I. Obaid, and A. Alnedawi. "Fatal Pedestrian Crashes at Intersections: Trend Mining using Association Rules". In Accident Analysis and Prevention, Vol. 160, 2021, p. 106306.

37. Omranian, E., H. Sharif, S. Dessouky, and J. Weissmann. "Exploring Rainfall Impacts on the Crash Risk on Texas Roadways: A Crash-Based Matched-Pairs Analysis Approach". In Accident Analysis and Prevention, Vol. 117, 2018, p. 10-20.

38. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805 [cs], 2019.

39. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. "Roberta: A robustly optimized bert pretraining approach". arXiv Preprint arXiv, 2019, 1907.11692.

40. Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz. "Transformers: State-of-the-Art Natural Language Processing". 2020.

41. Maiya, A. S. "ktrain: A low-code library for augmented machine learning". In The Journal of Machine Learning Research, Vol. 23, No. 1, 2022, pp. 7070–7075.

42. Smith, L. N. "Cyclical Learning Rates for Training Neural Networks (arXiv:1506.01186)". arXiv, 2017. https://doi.org/10.48550/arXiv.1506.01186

43. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need" . arXiv, 2017. https://arxiv.org/abs/1706.03762

44. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". p. 9.

45. Pennington, J., R. Socher, and C. Manning. "Glove: Global Vectors for Word Representation". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, Association for Computational Linguistics Doha, Qatar, 2014.

46. Taylor, W. L. "'Cloze Procedure': A New Tool for Measuring Readability". In Journalism Quarterly, Vol. 30, No. 4, 1953, pp. 415–433. https://doi.org/10.1177/107769905303000401.

47. Phan, S., G. Henter, Y. Miyao, and S. Satoh. "Consensus-Based Sequence Training for Video Captioning". arXiv, 2017. https://arxiv.org/abs/1712.09532

48. Lin, T., P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal Loss for Dense Object Detection". arXiv, 2018. https://arxiv.org/abs/1708.02002

49. Kingma, D. P., and J. Ba. Adam. "A Method for Stochastic Optimization". http://arxiv.org/abs/1412.6980. Accessed June 9, 2022.

50. Roboflow Docs. https://docs.roboflow.com/. Accessed October 12, 2023.

51. Zoph, B., E. Cubuk, G. Ghiasi, T. Lin, J. Shlens, and Q. Le. "Learning Data Augmentation Strategies for Object Detection". In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, pp. 566-583. Springer International Publishing.

52. Bochkovskiy, A., C. Wang, and H. Liao. "Yolov4: Optimal Speed and Accuracy of Object Detection". arXiv preprint arXiv:2004.10934, 2020.

53. Gu, X., T. Li, Y. Wang, L. Zhang, Y. Wang, and J. Yao, "Traffic Fatalities Irediction using Support Vector Machine with Hybrid Particle Swarm Optimization." In Journal of Algorithms Computing Technology, Vol. 12, No. 1, pp. 20–29, 2018, doi: 10.1177/1748301817729953.

54. Sun, B., and B. Park. “Route Choice Modeling with Support Vector Machine.” In Transportation Research Procedia, Vol. 25, pp. 1806–1814, 2017, doi: 10.1016/j.trpro.2017.05.151.

55. Li, X., D. Lord, Y. Zhang, and Y. Xie. “Predicting Motor Vehicle Crashes using Support Vector Machine Models.” In Accident Analysis and Prevention, Vol. 40, No. 4, 2008, pp. 1611–1618, doi: 10.1016/j.aap.2008.04.010.

56. Kuhn, M. “Caret Package”. In Journal of Statistical Software, Vol. 28, No. 5, 2008.

57. Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. “LEGAL-BERT: The Muppets Straight out of Law School”. arXiv preprint arXiv:2010.02559, 2020.

58. He, P., J. Gao, and W. Chen. “Debertav3: Improving DeBERTA using Electra-Style Pre-Training with Gradient-Disentangled Embedding Sharing”. arXiv preprint arXiv:2111.09543, 2021.

59. Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations”. arXiv preprint arXiv:1909.11942, 2019.

60. Conneau, A., K., Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, and V. Stoyanov. “Unsupervised Cross-Lingual Representation Learning at Scale”. arXiv preprint arXiv:1911.02116, 2019.

61. Le, M., S. Geedipally, K. Fitzpatrick, N. Johnson, and R. Avelar. Understanding Dallas District Pedestrian Safety Issues. Report No. TTI-2019-4, Texas A&M Transportation Institute, College Station, Texas, 2019.

62. Doesken, N. J., T. B. McKee, and C. Davey. “Climate Data Continuity–What Have we Learned from the ASOS Automated Surface Observing System”. In Proceedings of the 13th Conference on Applied Climatology, Portland, Oregon, 2002, pp. 13–16.

63. DeGaetano, A. T., W. Noon, and K. L. Eggleston. “Efficient Access to Climate Products using ACIS Web Services”. In Bulletin of the American Meteorological Society, Vol. 96, No. 2, 2015, pp. 173–180.

64. Sathiaraj, D., T.-o. Punkasem, F. Wang, and D. P. Seedah. “Data-Driven Analysis on the Effects of Extreme Weather Elements on Traffic Volume in Atlanta, GA, USA”. In Computers, Environment and Urban Systems, Vol. 72, 2018, pp. 212–220.

65. Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick. “Using Spatial Interpolation to Construct a Comprehensive Archive of Australian Climate Data”. In Environmental Modelling & Software, Vol. 16, No. 4, 2001, pp. 309–330.

66. The HDF Group, Hierarchical Data Format, version 5, 1997-2023, https://www.hdfgroup.org/HDF5/.

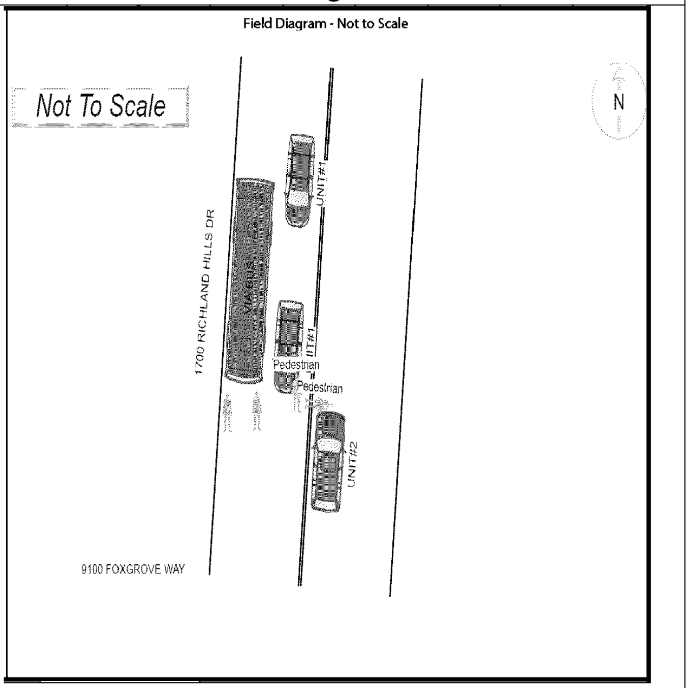
67. National Oceanographic and Atmospheric Administration, NOAA Solar Calculator, 2023, accessed on July 18, 2023.

68. Kidd B., M. Le, C. Poe, S. Joshua, and J. Short, Evaluating Recurring and Nonrecurring Congestion Impacts within Phoenix Metropolitan Region, Arizona. Transportation Research Board, Washington DC, 2012.

APPENDIX - CRASH REVIEW EXAMPLES

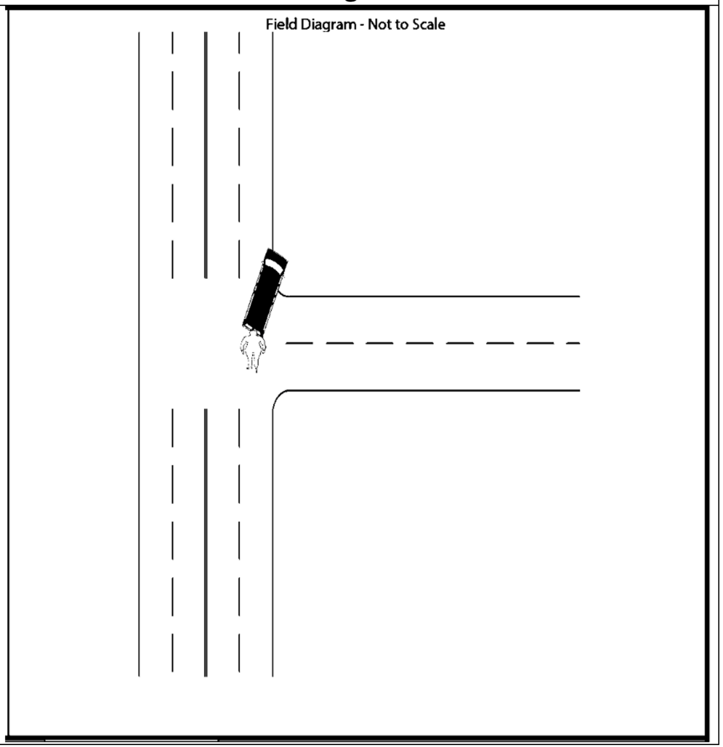
Crash 16631611 is an example of a good narrative and a good crash diagram. Researchers easily inferred this was a S-CR crash type and it was an ‘intended’ pedestrian (Table 31).

Table 31. Good Narrative and Good Crash Diagram Example.

Narrative	Diagram
DRIVER#1 STATED THAT PEDESTRIAN RAN IN FRONT OF HER VEHICLE AND SHE COULD NOT STOP PRIOR TO HITTING HIM. DRIVER#2 STATED THAT PEDESTRIAN GOT STRUCK BY UNIT#1 AND WAS THROWN INTO HER VEHICLE. DRIVER#2 STATED SHE WAS STOPPED AT TIME OF IMPACT. PEDESTRIAN STATED THAT HE RAN IN FRONT OF VIA BUS TO TRY AND CATCH ANOTHER BUS AND DID NOT CHECK TO SEE IF ANY TRAFFIC WAS COMING. WITNESS STATED PEDESTRIAN JUST RAN IN FRONT OF UNIT#1 AND UNIT#1 COULD NOT AVOID STRIKING PEDESTRIAN.	 Field Diagram - Not to Scale Not To Scale N UNIT#1 1700 RICHLAND HILLS DR VIA BUS UNIT#1 Pedestrian Pedestrian UNIT#2 9100 FOXGROVE WAY

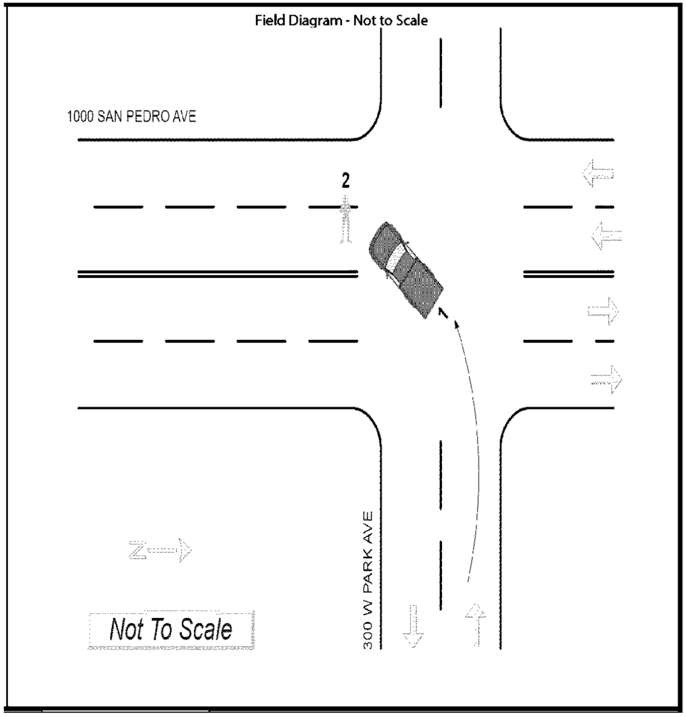
Crash 16730041 is an example of a bad narrative and a bad crash diagram. Researchers inferred this was a B-UN crash type because Unit 2 (pedestrian) movement was unknown. The diagram only showed unit 1 and unit 2 without any indication of direction of travel. It was assumed that Unit 2 was also the driver of the "another vehicle" from previous crash referenced in the narrative. Thus, it was concluded that this was a case of an 'unintended' pedestrian (Table 32).

Table 32. Bad Narrative and Bad Crash Diagram Example.

Narrative	Diagram
UNIT 1 WAS ATTEMPTING TO DISLODGE UNIT 1 FROM ANOTHER VEHICLE THAT IT WAS INVOLVED WITH IN A SEPARATE COLLISION. WHEN THE VEHICLE CAME DISLODGED, UNIT 2 WAS WALKING BEHIND UNIT 1. THE BACK RIGHT OF UNIT 1 STRUCK THE FRONT OF UNIT 2. UNIT 1 DRIVER SATID THAT HE TOLD UNIT 2 HE WAS ABOUT TO BACK THE VEHICLE UP.	

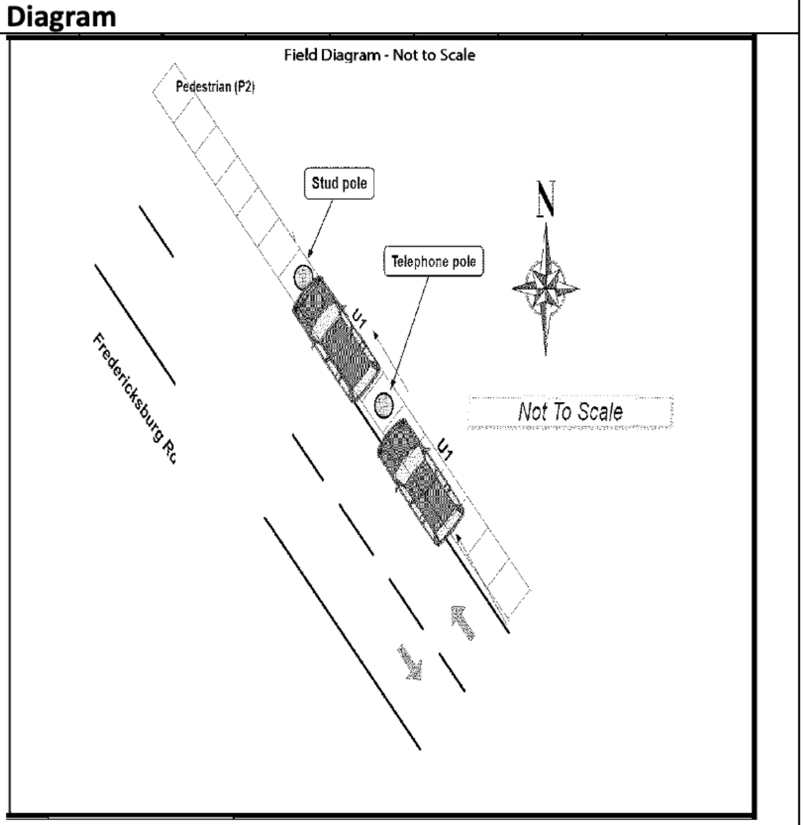
Crash 16686619 is an example of a “good” narrative but a “bad” diagram because diagram properly labeled the units involved but only shows the direction of Unit 1 (Table 33).

Table 33. Good Narrative and Bad Crash Diagram (First Example).

Narrative	**Diagram**
UNIT1 STOPPED AT THE INTERSECTION FACING WEST PREPARING TO TURN SOUTH ON 1000 SAN PEDRO AVE. UNIT2 ENTERED THE CROSSWALK WALKING WESTBOUND ON FOOT. UNIT1 PROCEEDED FORWARD AND INITIATED A LEFT TURN. UNIT1 DID NOT OBSERVE UNIT2 AND STRUCK UNIT2. UNIT2 STATED SHE STUBBLED AND WAS PUSHED BACK BUT DID NOT FALL TO THE GROUND. UNIT1 AT FAULT. UNIT1 DRIVER MADE CONTACT WITH UNIT2. UNIT2 STATED SHE WAS FINE. UNIT1 LEFT LOCATION. UNIT2 STATED SHE LATER BEGAN TO FEEL PAIN IN HER NECK AND BACK AND MOVED TO THE HOSPITAL FOR EXAMINATION. UNIT2 HAD NO FURTHER INFORMATION ON UNITI1.	 Field Diagram - Not to Scale 1000 SAN PEDRO AVE 2 N Not To Scale 300 W PARK AVE

Crash 16779339 is another example of a “good” narrative but a “bad” diagram. The narrative clearly described the situation and even mentions the keyword “debris” which is typical of the N-FC crash type. But it is difficult to tell where the pedestrian is or its direction of travel in the diagram (Table 34).

Table 34. Good Narrative and Bad Crash Diagram (Second Example).

Narrative	Diagram
Driver of U1 was traveling NW on Fredericksburg RdA. D1 lost control of U1 and veered off the roadway to the right of the street onto the sidewalk and struck a telephone pole (listed in property line) and a stud pole (listed in property line) directly after striking the first one. U1 spun after hitting both poles and ended up facing SE on Fredericksburg Rd. P2 stated he was walking SE on the sidewalk of Fredericksburg and observed U1 crash into poles. P2 stated he was struck by debris from the poles on his right leg. P2 was seen by EMS and had minor scrapes on his right leg. BWC/Coban available Lab results updated [retracted]	 Field Diagram - Not to Scale Pedestrian (P2) Stud pole Telephone pole N U1 U1 Fredericksburg Rd Not To Scale

Texas A&M
Transportation
Institute

tti.tamu.edu

Minh Le
TTI Research Engineer
Program Manager
Research & Implementation — Dallas
Texas A&M Transportation Institute
12700 Park Central, Suite 1000
Dallas, TX 75251
(972) 994-2212
M-Le@tti.tamu.edu