

INTRODUCTION

In the state of Maryland, the 2019-2023 5-year average of total crashes is over 107,000; the average number of injured people is 41,717. There are various costs associated with such fatalities and injuries: to the person, to the families, to the communities. Scholars have estimated that injuries and deaths caused by road accidents are projected to cost the world economy USD 1.8 trillion from 2015 to 2030.

In recent years, machine learning models like Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting, and Neural Networks have been effective at capturing non-linear relationships and complex interactions among influencing factors. Scholars have applied RF, Extreme Gradient Boosting (XGBoost) and logistic regression. The explicit implementation of feature analysis shows more detailed understanding of the relative importance of different features when predicting crash injury severity (1-3).

Scholars applying Natural Language Processing (NLP) to study crash reports and narratives, and image analysis of crash diagrams (publicly available examples of crash diagrams):

Crash Narrative Example (CISS)

V1 was traveling east in lane two of a three lane roadway physically divided by a concrete barrier. V2 was traveling east on the same roadway in lane one. V1 departed its original lane of travel to the right. The right of V1 contacted the left of V2. Both vehicles came to a final rest facing east on the right.

Crash Diagram Example

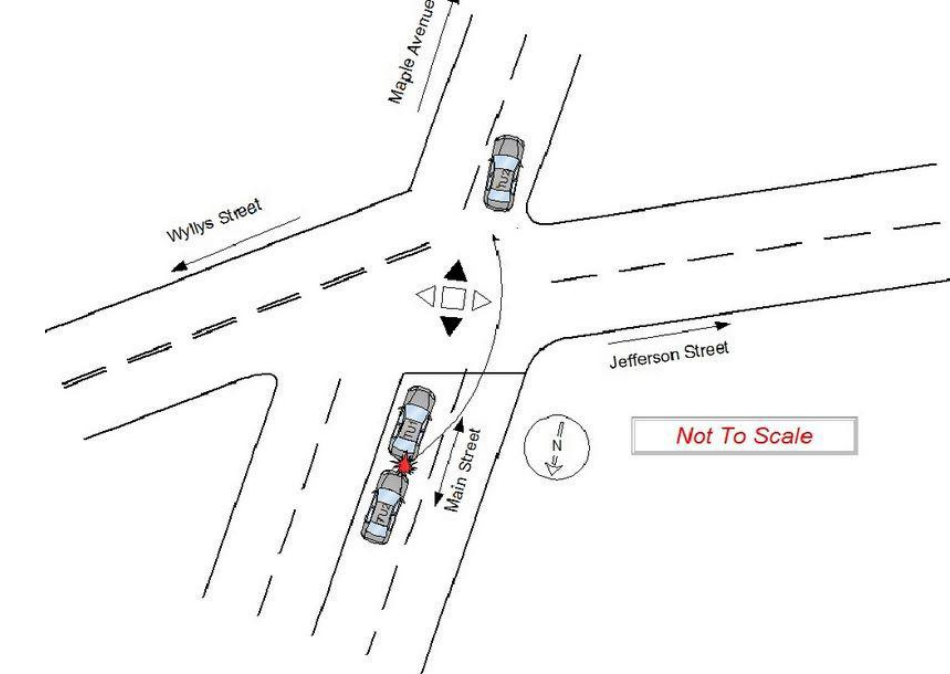


Photo credit: Connecticut Transportation Safety Research Center

METHODS AND DATA

Data: MD Crash Reports for 2016-2022, N=47,000, the analysis reported here are based on a smaller representative random sample (n=1000) due to computational capacity and time limitations. Occupants in cars, SUVs, light trucks; non-traffic crashes were excluded.

Type of Crash	Frequency	Percent
Fatal	357	0.8
Injury	10574	24.5
Property damage	32303	74.7
Total	43234	100.0

AI Methods and Models: (1) Natural Language Processing (NLP) for crash narratives analysis. (2) Structural Topic Modeling (STP) for crash narrative topics extraction. (3) Convolutional Neural Networks (CNN) for image analysis and features extractions for crash diagrams. (4) Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for reducing dimensions for crash diagrams. (5) ML models: Classification and Regression Tree (CART), Random Forests, XGBoost and SHAP analysis for factor importance and predictive modelling. Data: crash characteristics, crash narrative latent topics, crash diagrams extracted main features.

Latent Topics Description: (1) Dominant characteristic - vehicles making improper turns, changing lanes or at intersections. (2) Dominant characteristic - one vehicle involved, sometimes in interaction with a tractor trailer or a truck. (3) Dominant characteristic: vehicles either stopped in traffic, parked and/or bus involved. (4) Dominant characteristic: vehicles travelling on major highways, I-95, I-695, route 80, etc. (5) Dominant characteristic: specifics found in the narrative about vehicle occupants receiving, not receiving or refusing medical attention.

RESULTS

I. Crash Narrative Analysis

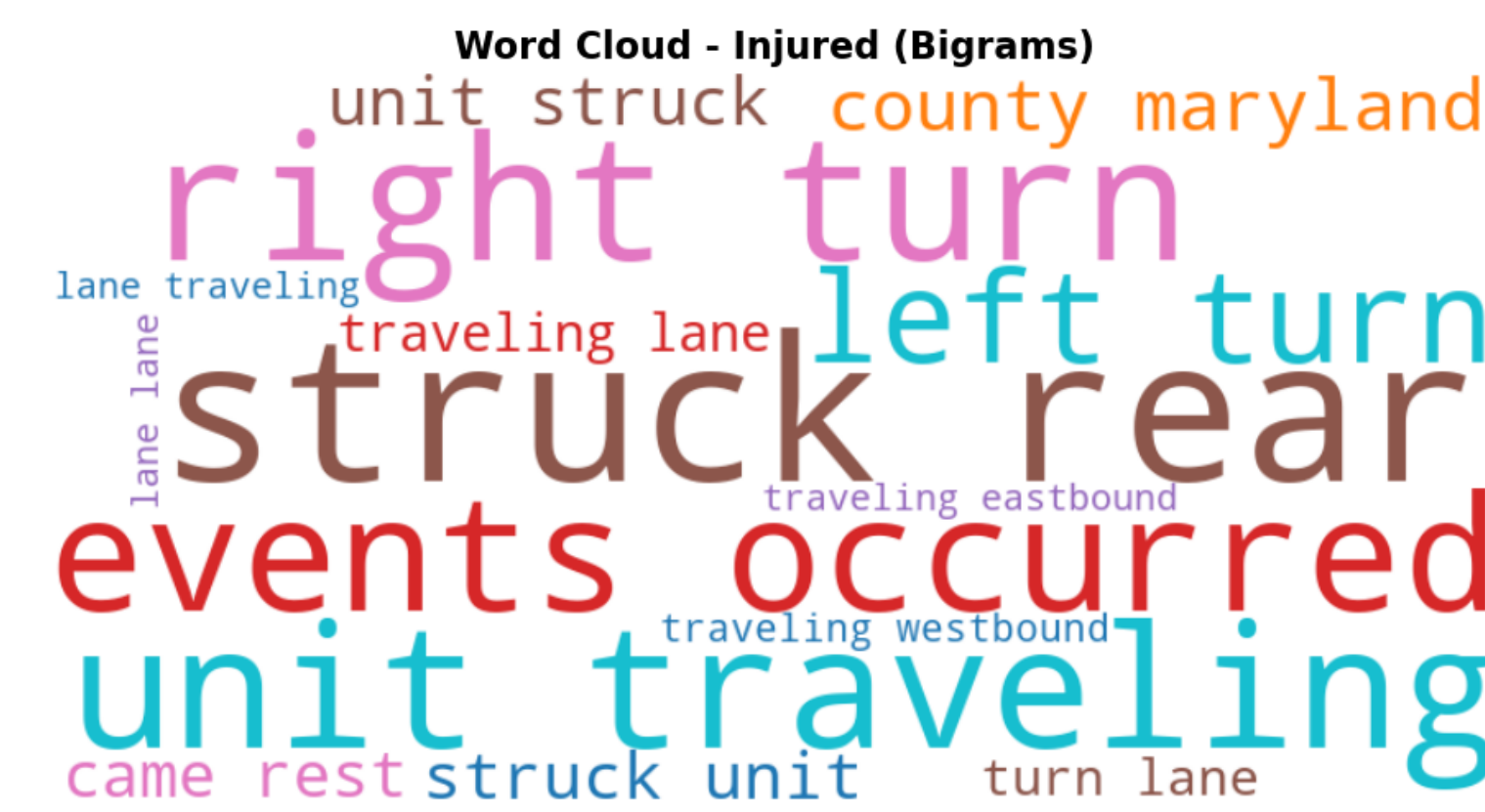


Figure 1. Distinctive two-word patterns (bi-grams) for injury crashes

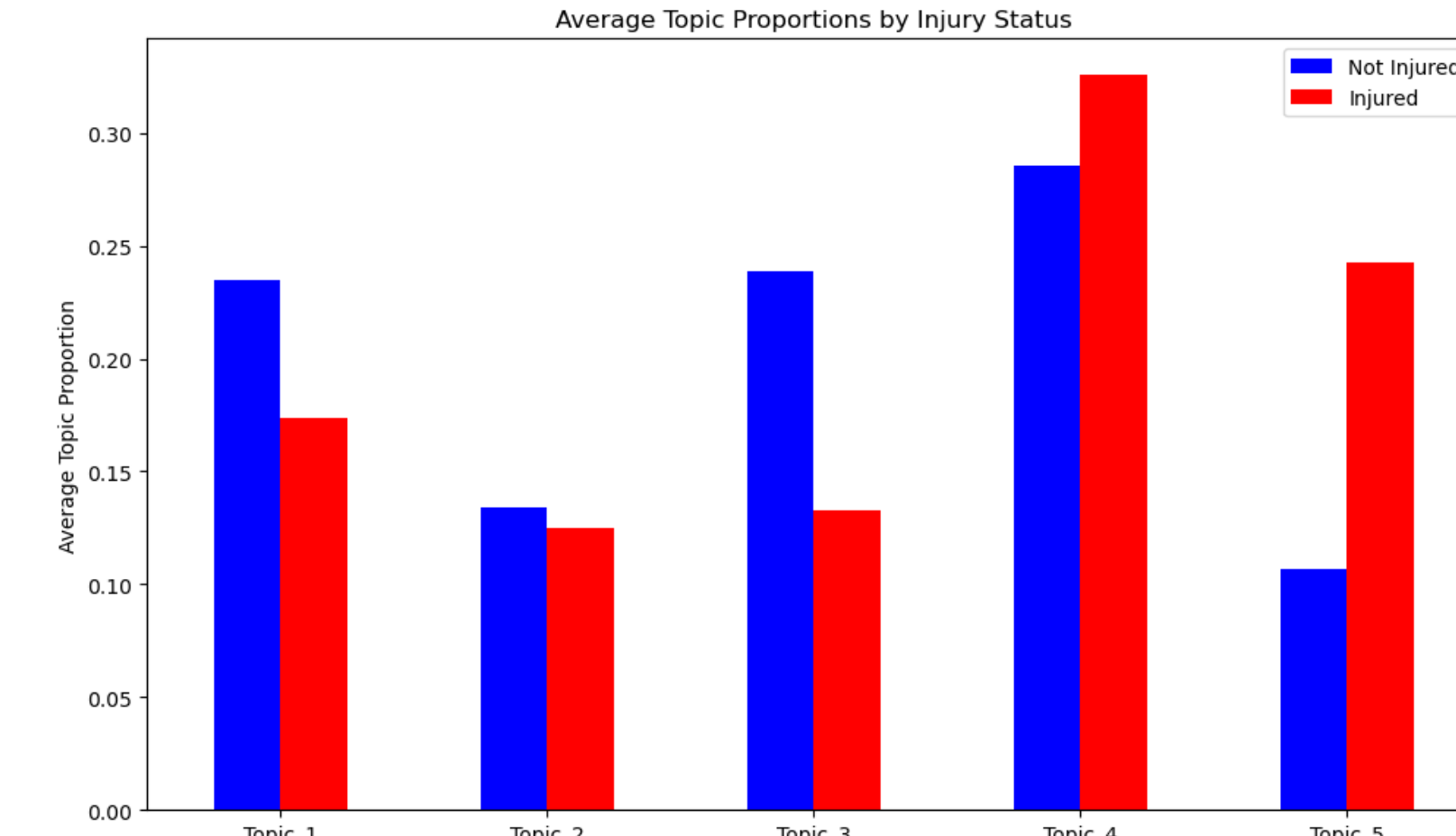


Figure 2. Crash Narratives: AI Extracted Latent Topics, Comparison Not Injured vs. Injured

Rank	Crash Characteristics Only AUC=0.54	Crash Characteristics & AI Crash Narratives Topics AUC=0.7
1	Time of the crash (0-23 hour)	AI Crash Narrative Topic 5
2	Road: two way, not divided	AI Crash Narrative Topic 2
3	Month of the crash	AI Crash Narrative Topic 3
4	Day of the week of the crash	Road: two way, not divided
5	Collision: same direction, side swipe	Hit and Run
6	Hit and run	Month of the crash
7	Number of lanes	AI Crash Narrative Topic 1
8	Collision: rear end	Collision: rear end
9	Intersection	Day of the week of the crash
10	Traffic control (Y/N)	Number of lanes

Figure 3. ML Models: Most Important Factors for Injury/Non-Injury Crashes; Sample N=1000 crash reports and narratives.

II. Crash Diagram Analysis

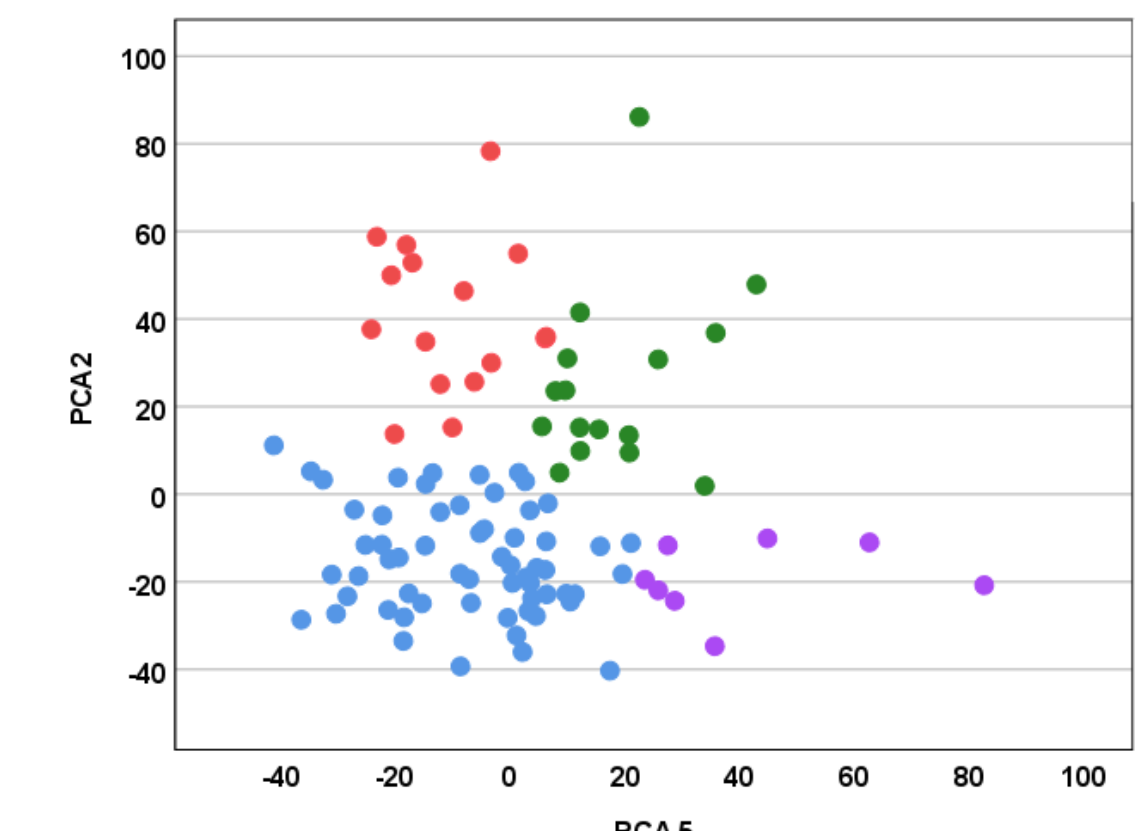


Figure 4. Analysis of Crash Diagrams. Based on random sample of 100 crash diagrams. Deep Learning: Convolutional Neural Networks (CNN); Four Clusters.

Rank	Crash Characteristics Only AUC=0.5	Crash Characteristics & AI Crash Diagram Features AUC=0.6
1	Time of the crash (0-23 hour)	AI Crash Diagram Feature 5 (PCA)
2	Road: two-way, not divided	Road: two-way, not divided
3	Daylight (Y/N)	AI Crash Diagram Feature 2 (PCA)
4	Intersection (Y/N)	Time of the crash (0-23 hour)
5	Collision: rear end	Daylight (Y/N)
6	Number of lanes	AI Crash Diagram Feature 3 (PCA)
7	Traffic control (Y/N)	Intersection (Y/N)
8	Collision: same direction, side swipe	Collision: rear end
9	Month of the crash	Traffic control (Y/N)
10	Day of the week of the crash	Number of lanes

Figure 5. ML Models: Most Important Factors for Injury/Non-Injury Crashes. Sample N=100 crash reports with crash diagrams.

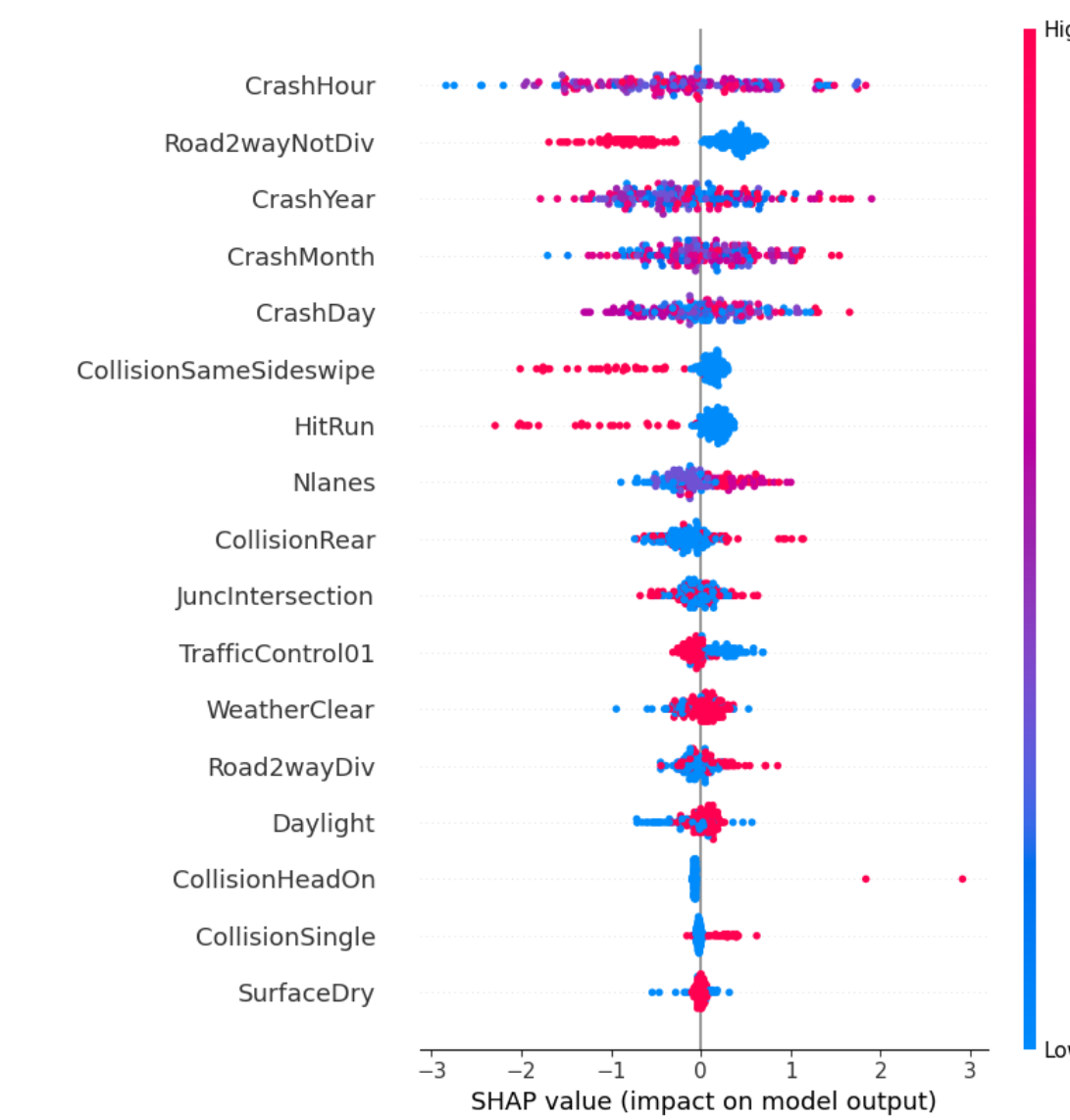


Figure 6. XGBoost SHAP analysis: Shows how each feature shifts predictions of Injury vs Non-injury crash.

III. Predictive Models

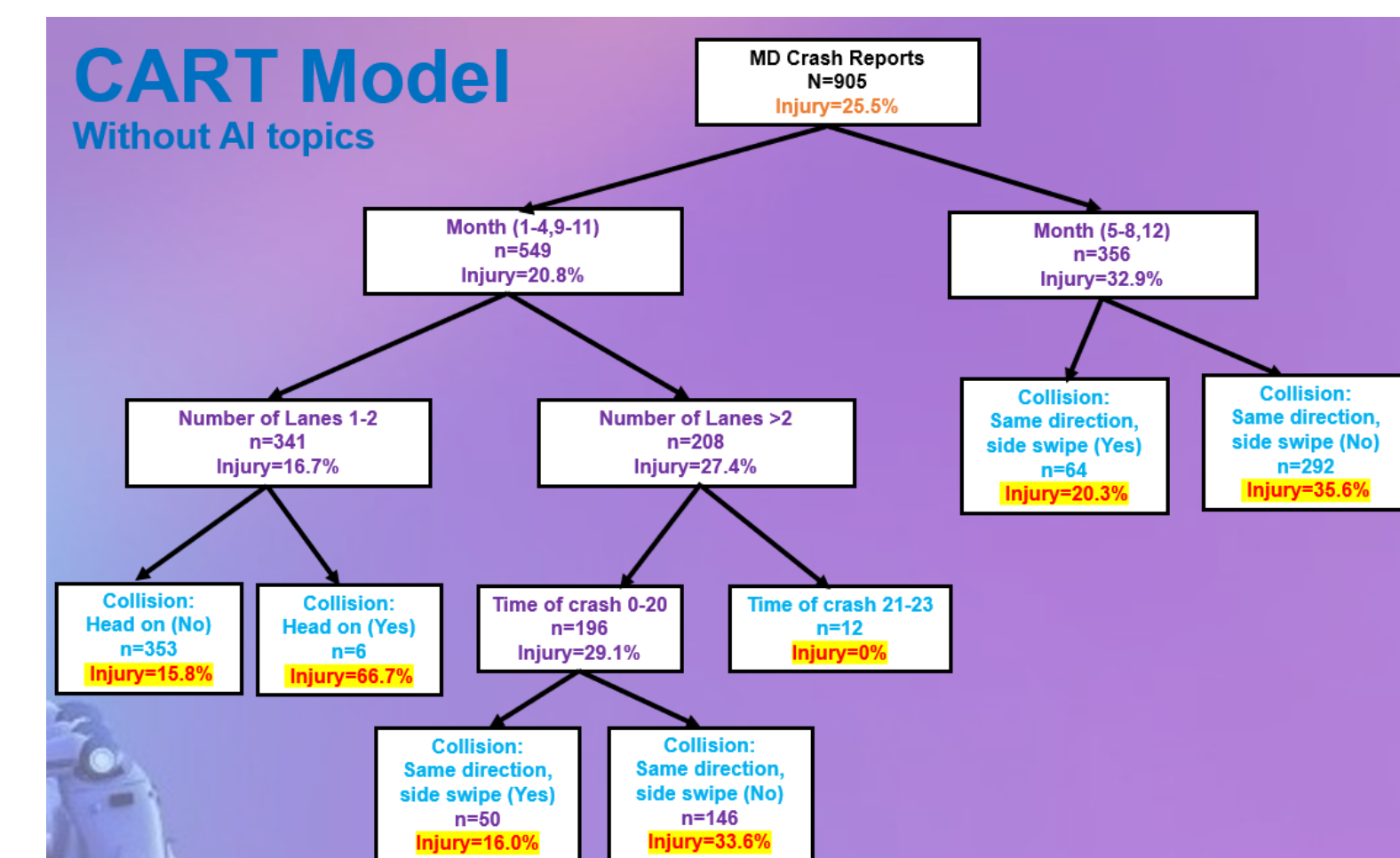


Figure 7. Cart Models, no AI Topics

#	Injury (%)	Injury Level	Crash Characteristics
1	0	Low	Month (1-4,9-11), Number of lanes>2, Time of crash 21-23.
2	15.8	Low	Month (1-4,9-11), Number of lanes 1-2, Collision head on (No)
3	16.0	Low	Month (1-4,9-11), Number of lanes >2, Time of crash (0-20), Collision same direction, side swipe (Yes)
4	20.3	Medium	Month (5-8,12), Collision same direction, side swipe (Yes)
5	33.6	High	Month (1-4,9-11), Number of lanes >2, Time of crash (0-20), Collision same direction, side swipe (No)
6	35.6	High	Month (5-8,12), Collision same direction, side swipe (No)
7	66.7	High	Month (1-4,9-11), Number of lanes 1-2, Collision head on (Yes)

Figure 8. CART Model, no AI topics, Injury Groups

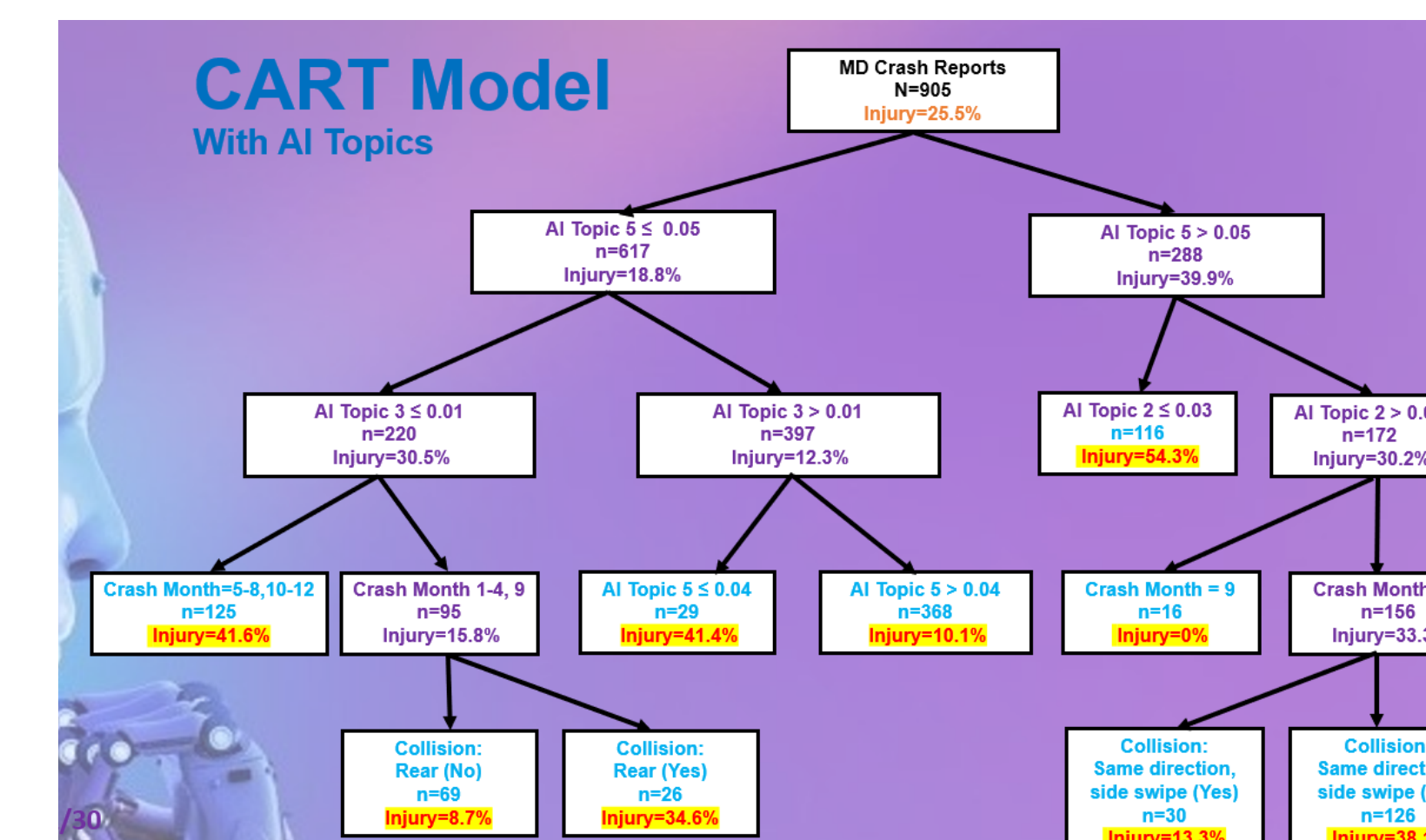


Figure 9. CART Model, with AI Topics

#	Injury (%)	Injury Level	Crash Characteristics and Crash Narratives AI Topics
1	0	Low	AI Topic 5 > 0.05, AI Topic 2 > 0.03, Crash Month = 9
2	8.7	Low	AI Topic 5 ≤ 0.05, AI Topic 3 ≤ 0.01, Crash Month=1-4, 9, Collision rear (No)
3	10.1	Low	AI 0.04 < Topic 5 ≤ 0.05, AI Topic 3 > 0.01
4	13.3	Low	AI Topic 5 > 0.05, AI Topic 2 > 0.03, Crash Month ≠ 9, Collision same direction side swipe (Yes)
5	34.6	High	AI Topic 5 ≤ 0.05, AI Topic 3 ≤ 0.01, Crash Month=1-4, 9, Collision rear (Yes)
6	38.1	High	AI Topic 5 > 0.05, AI Topic 2 > 0.03, Crash Month ≠ 9, Collision same direction side swipe (No)
7	41.4	High	AI Topic 5 ≤ 0.04, AI Topic 3 > 0.01
8	41.6	High	AI Topic 5 ≤ 0.05, AI Topic 3 ≤ 0.01, Crash Month=5-8, 10-12
9	54.3	High	AI Topic 5 > 0.05, AI Topic 2 ≤ 0.03

Figure 10. CART Model, with AI topics, Injury Groups

CONCLUSION

- The new techniques associated with Natural Language Processing (NLP) give us the opportunity to discover new relationships between crash characteristics and injury/non-injury outcomes.
- Crash Narrative Analysis:** We discovered **five latent topics** associated with the crash narratives, which significantly improve the predictive models.
- Crash Diagram Analysis:** We discovered **five latent features** associated with the crash diagrams, which improve the predictive models and help us distinguish between four crash clusters. Furthermore, **four clusters** were established with specific implications
- Cluster 1:** Dominant characteristic – vehicles travelling in the same general direction, e.g. side-swipe, no head-on contact.
- Cluster 2:** Dominant characteristic – vehicles at or approaching an intersection, including a railroad, and/or involving a bus.
- Cluster 3:** Dominant characteristic - vehicles traveling on major highways, I-95, I-695, etc.
- Cluster 4:** Dominant characteristic – crashes involving at least one large vehicle: e.g., a tractor-trailer or a truck.

FUTURE DIRECTIONS

- The most recent methods and models based on Artificial Intelligence, and specifically NLP, are very promising in discovering and extracting non-linear data patterns. They can be utilized to increase road safety.
- MHSO could work with NSC to provide feedback and insights about the newly discovered crash patterns.
- Natural Language Processing can be utilized to improve the written narratives submitted by law enforcement.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of Maryland Highway Safety Office (MHSO). All opinions expressed in this report are solely of the authors and not necessarily of MHSO.

We would like to thank Maryland State Police (MSP) for the access to the crash data.

LITERATURE CITED

- Ekanem, I. (2025) Analysis of Road Traffic Accident Using AI Techniques. Open Journal of Safety Science and Technology, 15, 36-56.
- Qawasmeh B, Oh J-S, Kwizgile V. Investigating Injury Outcomes of Horse-and-Buggy Crashes in Rural Michigan by Mining Crash Reports Using NLP and CNN Algorithms. Safety. 2025; 11(1):1
- Lee, S.-S., Cha, S.-M., Ko, B., et al. 2023, IEEE Access, Extracting Fallen Objects on the Road From Accident Reports Using a Natural Language Processing Model-Based Approach, 11, 139521.