Interpretable Feature Selection for Truck Collision Injury Severity: A SHAP-RFE Approach

Sana Chodri
sana.chodri@torontomu.ca
Toronto Metropolitan University
Toronto, Ontario, Canada

Hossein Zolfagharinia
Toronto Metropolitan University
Toronto, Ontario, Canada

Aida Haghighi
Toronto Metropolitan University
Toronto, Ontario, Canada

Abstract

Truck-involved collisions pose a significant safety and operational risk within supply chains, often resulting in costly disruptions, injuries, and delays. Accurate and interpretable prediction of injury severity is critical for supporting proactive safety interventions and risk mitigation strategies. This study presents a SHAP-guided Recursive Feature Elimination (SHAP-RFE) framework for identifying the most informative features related to injury severity in truck crashes, using data from the 2022 Fatality Analysis Reporting System (FARS).

We compare SHAP-RFE against two benchmark feature selection methods: Principal Component Analysis (PCA) and a literature-informed feature set synthesized from 58 prior studies. Our approach achieves the highest adjusted macro F1-score, while selecting a compact set of 26 interpretable features. Notably, 20 of these overlap with domain-validated risk factors, confirming strong alignment with existing research.

The results highlight SHAP-RFE's ability to balance performance and interpretability in imbalanced multiclass classification tasks. This interpretable framework offers practical value for transportation safety planners and logistics decision-makers seeking to reduce crash impact and enhance supply chain resilience.

CCS Concepts

• **Applied computing → Transportation**; *Supply chain management*; • **Computing methodologies → Dimensionality reduction and manifold learning**; *Classification and regression trees*; *Supervised learning by classification.*

Keywords

interpretable machine learning, feature selection, SHAP, truck collisions, multiclass classification, injury severity prediction

ACM Reference Format:
Sana Chodri, Hossein Zolfagharinia, and Aida Haghighi. 2025. Interpretable Feature Selection for Truck Collision Injury Severity: A SHAP-RFE Approach. In *Proceedings of The 4th Workshop on Artificial Intelligence for Supply Chain (AI4SupplyChain @ KDD 2025).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

AI4SupplyChain @ KDD 2025, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/2025/08
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Truck-involved collisions pose a significant safety and operational risk within logistics and supply chain management, affecting transportation reliability, insurance liability, and service continuity. In the United States alone, trucks transported over 11.84 billion tons of freight in 2019, with commercial vehicles accounting for nearly 30% of all traffic-related fatalities [1,2]. Globally, the freight trucking market was valued at $2.2 trillion in 2022 and is projected to reach $3.4 trillion by 2030, underscoring the high economic stakes associated with truck safety and efficiency [3]. As freight demand rises, especially with the growth of e-commerce and just-in-time logistics—understanding and mitigating truck collision severity has become an urgent research and management priority.

Although machine learning (ML) has been increasingly applied to forecasting, optimization, and risk detection tasks in logistics, its application to injury severity prediction in truck-involved crashes remains limited. Moreover, many existing models in this domain emphasize predictive performance over interpretability, reducing their usefulness in real-world, high-stakes decision-making contexts. In transportation safety—where legal, operational, and policy implications are substantial—interpretable ML methods are essential for enabling actionable insights and fostering trust among practitioners and regulators.

To address this gap, we adapt and evaluate a SHAP-guided Recursive Feature Elimination (SHAP-RFE) framework for interpretable feature selection in multiclass injury severity prediction. Originally introduced by Huang et al. (2024) for binary classification of driver mental states using physiological data, SHAP-RFE combines SHapley Additive Explanations (SHAP)—a game-theoretic method for feature attribution—with recursive elimination to identify compact, high-importance feature subsets [4]. Unlike traditional filter- or wrapper-based methods, this approach seeks to balance predictive performance with transparency, making it well-suited for high-stakes, safety-critical applications such as transportation risk modeling.

Feature selection plays a pivotal role in injury severity modeling by reducing dimensionality, enhancing generalizability, and improving model transparency—particularly when used to guide safety interventions or policy decisions [5]. In the context of commercial trucking, identifying a focused, interpretable set of risk factors can directly inform targeted safety strategies and resource allocation.

This study asks: *Can SHAP-RFE provide a feature selection method that balances interpretability and predictive performance, outperforming standard approaches such as Principal Component Analysis (PCA) and literature-informed variables—in truck-involved injury severity classification?*

To investigate this, we apply SHAP-RFE to real-world crash data from the Fatality Analysis Reporting System, 2022 and benchmark it against two baselines: (i) PCA, an unsupervised technique that emphasizes variance rather than interpretability, and (ii) a literature-informed feature set derived from an extensive synthesis of 58 prior studies. This three-way comparison enables a comprehensive evaluation of SHAP-RFE's utility as a methodologically grounded and operationally relevant feature selection framework in commercial vehicle safety analytics.

2 Related Work

Machine learning (ML) techniques have seen increasing adoption in traffic safety research, yet their application to truck-involved crash injury severity prediction remains limited. The vast majority of prior studies in this domain rely on traditional statistical models—including multinomial logit, ordered probit, and random parameters logit—to examine crash-related outcomes [6,7,8]. While these models offer interpretability and simplicity, they impose restrictive assumptions about linearity, independence among predictors, and fixed functional forms, which limit their ability to capture the nonlinear and high-dimensional interactions common in real-world crash data.

A smaller but growing body of work has explored machine learning (ML) methods, particularly ensemble models like gradient boosting and deep neural networks, for crash severity prediction. These models have demonstrated strong predictive performance and are especially capable of capturing nonlinear relationships and complex feature interactions [9,10]. However, they often prioritize predictive accuracy over interpretability, and most studies use all available features without applying formal feature selection. When feature importance is reported, it is typically done post hoc and based on model-specific heuristics, limiting the transparency and actionability of findings.

To improve interpretability, some recent studies have adopted SHapley Additive Explanations (SHAP), a model-agnostic method for attributing feature influence. For example, Yang et al. (2021) and Abdulrashid et al. (2024) apply SHAP to analyze feature contributions in crash severity models. However, in both cases, SHAP is used only descriptively after model training. It is not integrated into a systematic feature selection pipeline—limiting its potential to guide the development of compact, interpretable models suitable for operational use [10,11].

Our study addresses these gaps by proposing a SHAP-guided Recursive Feature Elimination (SHAP-RFE) framework that combines transparency and predictive rigor. By integrating SHAP within a recursive feature selection loop—and benchmarking it against PCA and a literature-informed set grounded in over 58 empirical studies—we contribute a methodologically grounded, operationally relevant approach for risk factor identification in truck-involved crash severity prediction.

3 Methodology

This study follows a three-phase methodological pipeline to evaluate the effectiveness of SHAP-guided Recursive Feature Elimination (SHAP-RFE) for interpretable feature selection in truck-involved crash severity prediction. First, we prepare the dataset from FARS 2022 and define the injury severity target variable. Second, SHAP-RFE is applied using a CatBoost classifier to iteratively eliminate low-importance features based on multiclass SHAP aggregation. Finally, we benchmark SHAP-RFE against two baseline approaches: Principal Component Analysis (PCA) and literature-informed feature sets. All three subsets are evaluated using the same model architecture and macro F1-score.

3.1 Dataset and Preprocessing

We used the Fatality Analysis Reporting System (FARS) 2022, a nationally maintained database by the National Highway Traffic Safety Administration (NHTSA), which provides police-reported fatal crash records across the United States. Five structured files—Accident, Vehicle, Person, Distract, and DriverRF—were merged using unique case and vehicle identifiers to create an event-level dataset containing vehicle configuration, crash details, driver demographics, and injury outcomes.

To focus the analysis on truck-involved collisions, the data was filtered using vehicle body type and configuration codes to isolate commercial trucks. The injury severity target variable was derived from the KABCO scale and grouped into three classes: no injury (Class 0), minor injury (Class 1), and major/fatal injury (Class 2), consistent with prior injury severity modeling studies [12,13]. A binary fault indicator was constructed by examining crash-level and vehicle-level contributing factors to identify whether the commercial truck driver or other vehicle was at fault.

Preprocessing involved removing features with high missingness, low variance, or strong multicollinearity, followed by imputation using mean, median, or mode based on variable type. Categorical variables were one-hot encoded, with high-cardinality features manually grouped into broader, interpretable categories to reduce dimensionality. An 80–20 train-test split was applied, and SMOTE was used on the training set to address class imbalance [14].

The final dataset consisted of 4,098 observations and 74 features, reduced from an initial 5,085 records and 109 variables following filtering and preprocessing.

3.2 SHAP-RFE Feature Selection

To identify a compact and interpretable subset of features for multiclass injury severity prediction, we implemented SHAP-guided Recursive Feature Elimination (SHAP-RFE). This approach combines model-agnostic Shapley-based feature attribution with recursive feature removal to optimize both interpretability and predictive utility.

SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee [15], is an explainable AI technique grounded in cooperative game theory. It attributes a model's output to its input features by computing each feature's marginal contribution across all possible subsets. For a given instance j and feature i, the SHAP value $\phi_i^{(j)}$ represents how much feature i contributes to the model prediction for that instance.

Unlike traditional feature importance scores—such as Gini gain or split count in decision trees—SHAP satisfies desirable theoretical properties including local accuracy, consistency, and missingness. These properties ensure that feature attributions are additive, stable across models, and faithful to the actual prediction function.

Moreover, while tree-based importance scores can be biased toward high-cardinality features or unstable under data perturbations [16], SHAP provides more robust, interpretable explanations at both global and individual levels.

We used CatBoost, a gradient boosting algorithm optimized for categorical features and tabular data, to generate SHAP values via its internal TreeSHAP implementation. Given the multiclass nature of our task (three injury severity levels), SHAP produces a separate attribution vector per class. To obtain a unified importance ranking across all classes, we aggregated SHAP values using the following formula:

$$\text{Importance}(x_i) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K}\left|\phi_{i,k}^{(j)}\right|$$

where $\phi_{i,k}^{(j)}$ is the SHAP value of feature x_i for class k and instance j, and $K = 3$ is the number of injury severity classes. This aggregation ensures that features influential across multiple classes are appropriately weighted in the ranking.

Recursive elimination was performed by iteratively removing the feature with the lowest aggregated SHAP importance, retraining the model, and re-evaluating performance at each step. The recursive elimination process continued until the performance curve exhibited an elbow, indicating that further feature removal would lead to a sharp degradation in model performance. To determine this point, we plotted the adjusted macro F1-score at each iteration and identified the inflection point where gains plateaued and predictive accuracy began to decline.

We selected the adjusted macro F1-score as our evaluation metric, as it offers a balanced view of performance in imbalanced multiclass settings. Unlike standard accuracy, which can be biased toward the majority class, the adjusted macro F1-score accounts for the relative frequency of each class, weighting each F1-score accordingly. For K classes, the adjusted macro F1 is computed as:

$$\text{Adjusted Macro F1} = \sum_{k=1}^{K}\frac{n_k}{N}\cdot F1_k$$

where $F1_k$ is the F1-score for class k, n_k is the number of true instances in class k, and N is the total number of instances.

This SHAP-RFE approach enables interpretable and theoretically grounded feature selection tailored to the needs of safety-critical, supply chain relevant prediction tasks.

3.3 Baseline Feature Selection Methods

To benchmark the effectiveness of SHAP-RFE, we compared it against two widely used baseline feature selection strategies: Principal Component Analysis (PCA) and a literature-informed feature set derived from prior studies on truck-involved crash severity.

PCA serves as an unsupervised dimensionality reduction method that transforms the original feature space into a set of orthogonal components that capture the highest variance. We applied PCA to the full dataset (excluding categorical variables) and retained the top k components, where k matched the number of features selected by SHAP-RFE. Although PCA is frequently used to reduce dimensionality and potentially improve performance, it sacrifices interpretability, as the resulting components are linear combinations of original features without direct semantic meaning.

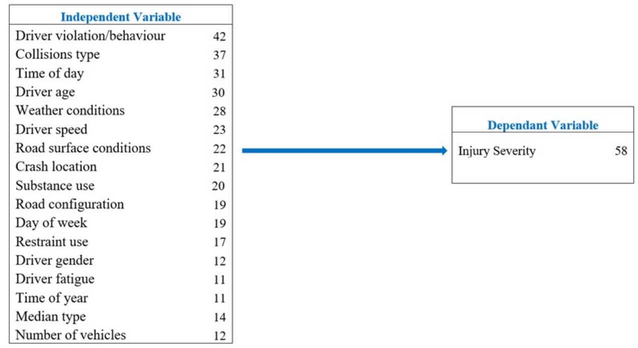

Figure 1: Classification of key variables from prior studies on truck-involved injury severity.

The literature-informed feature set was constructed by reviewing variables consistently identified as significant in 58 prior empirical studies on truck crash severity from the past decade. These included driver age, time of crash, vehicle configuration, crash type, and environmental conditions, among others. This set represents a theory-driven, domain-grounded approach based on expert understanding rather than algorithmic inference. Figure 1 provides a structured overview of the most commonly studied variables across the literature, organized by their roles as independent factors influencing injury severity.

All three feature sets—SHAP-RFE, PCA, and literature-informed—were evaluated using the same CatBoost classifier and macro F1-score. This triangulated comparison provides a balanced assessment of SHAP-RFE's ability to preserve predictive performance while improving interpretability over both data-driven and theory-based alternatives.

4 Results

The goal of this study is to evaluate the effectiveness of SHAP-guided Recursive Feature Elimination (SHAP-RFE) in selecting a compact and interpretable subset of features for multiclass injury severity prediction in truck-involved crashes. We compare SHAP-RFE against two baseline methods—Principal Component Analysis (PCA) and a literature-informed feature set using the same CatBoost model architecture and training pipeline.

4.1 Injury Severity Distribution and Class Imbalance

Before applying any sampling or modeling strategies, we examined the class distribution of the injury severity target variable. As shown in Figure 2, the dataset was highly imbalanced, with the majority of cases classified as *No Injury (0)*, while *Minor Injury (1)* and *Major Injury (2)* were significantly underrepresented. This imbalance posed a challenge for multiclass classification, as conventional accuracy metrics tend to overrepresent the majority class.

To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data, generating synthetic examples for the minority classes to equalize class distribution. The resulting class balance is illustrated in Figure 3, showing an even

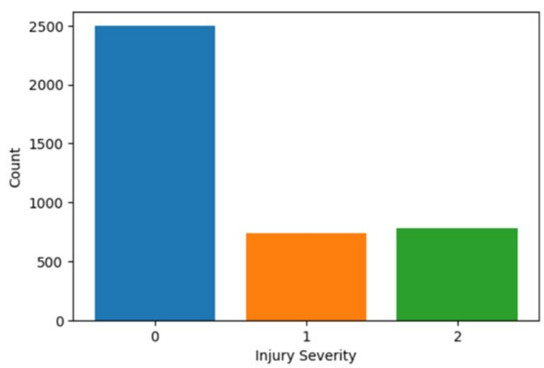

Figure 2: Class distribution of injury severity before applying SMOTE.

representation across all three severity levels. This balancing step ensured fairer model training and improved the reliability of evaluation metrics.

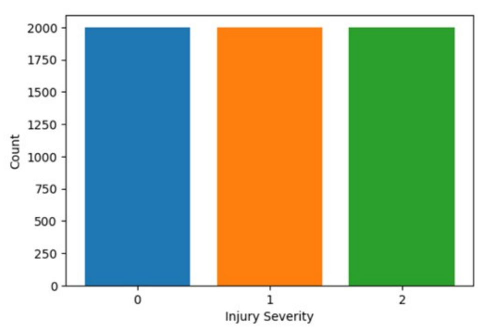

Figure 3: Class distribution of injury severity after applying SMOTE.

As previously mentioned, we used the *adjusted macro F1-score* to assess model performance. This metric offers a balanced view across classes by accounting for class frequency and averaging per-class F1-scores accordingly. It is particularly appropriate for imbalanced multiclass settings like ours.

4.2 Feature Performance Comparison

To assess the effectiveness of SHAP-guided Recursive Feature Elimination (SHAP-RFE), we compared it against two baseline feature selection methods: Principal Component Analysis (PCA) and a literature-informed feature set derived from prior studies on truck-involved crash severity.

All three feature sets were evaluated using the same CatBoost model architecture, trained on the SMOTE-balanced dataset, and assessed using the adjusted macro F1-score. As shown in Table 1, SHAP-RFE achieved the highest performance, with an adjusted macro F1-score of 0.45, followed by PCA (0.41) and the literature-informed features (0.32).

Table 1: Comparison of feature selection methods based on adjusted macro F1-score and interpretability.

Method	# Features	Adjusted Macro F1
SHAP-RFE	27	0.45
PCA	128	0.41
Literature-informed	44	0.32

These results suggest that SHAP-RFE effectively balances predictive performance and interpretability. While PCA provided moderately strong performance, its components are not directly interpretable, limiting its usefulness for decision-makers. The literature-informed feature set, although rooted in prior domain knowledge, underperformed relative to data-driven methods, highlighting the value of adaptive, model-aware selection strategies.

4.3 Overlap with Literature-Informed Features

To evaluate the domain relevance of the SHAP-RFE–selected features, we compared them against a literature-informed set derived from 58 prior studies on truck-involved crash severity. As shown in Table 2, 20 out of the 26 SHAP-RFE features overlapped with variables previously identified in the literature, including well-established risk factors such as driver age, crash type, vehicle age, time of day, and collision impact point [17,18].

This substantial overlap confirms that SHAP-RFE not only identifies statistically relevant features but also aligns strongly with existing domain knowledge. At the same time, the method uncovered five features not commonly reported in prior studies—such as driver home state, license restrictions, truck fault, driver weight, and National Highway System (NHS) route presence. These may reflect emerging or context-specific factors that warrant further investigation and validation in future studies.

Overall, the overlap analysis highlights SHAP-RFE's strength in bridging data-driven insights with theory-driven relevance. It demonstrates that interpretable machine learning methods can effectively surface both core and novel predictors, improving the potential for generalizability and policy impact.

5 Discussions

This study proposed and evaluated a SHAP-guided Recursive Feature Elimination (SHAP-RFE) framework for interpretable feature selection in the context of truck-involved crash injury severity. Compared to Principal Component Analysis (PCA) and a literature-informed feature set, SHAP-RFE achieved the highest adjusted macro F1-score (0.45) while selecting only 26 features. This demonstrates its ability to balance predictive performance with interpretability—an essential consideration in safety-critical applications.

5.1 Practical Contributions

While feature selection is a critical component of machine learning workflows, many prior studies on injury severity have relied on traditional statistical models or applied ML without robust feature selection [19,20]. These approaches often overlook variable redundancy, interaction effects, and lack interpretability. By embedding

Table 2: Comparison of SHAP-RFE Selected Features with Literature-Informed Features

SHAP-RFE Feature	Found in Literature
Driver Age	✓
Time of Day	✓
Crash Type	✓
Only Trucks	
Cargo Type	✓
Highway Type	✓
Number of Vehicles	✓
Crash Region	✓
Other Vehicle Fault	
Season	✓
Collision Impact Point	✓
First Harmful Event	✓
License Endorsement Status	
Rural/Urban	✓
Posted Speed Limit	✓
Driver Home State	✓
Driver Weight	✓
License Restrictions	
Number of Trailers	✓
Vehicle age	✓
Pre-crash Event	✓
Driver Height	
Trafficway Type	✓
National Highway System Route	
Truck Fault	✓
Manner of Collision	✓

SHAP values into an iterative elimination process, this study introduces a data-driven yet transparent feature selection method tailored for logistics and fleet risk modeling, an area where the balance between interpretability and predictive performance is often missing.

5.2 Implications for Supply Chain and Safety

Truck collisions and resulting injury severity carry significant consequences for logistics operations, including delivery delays, increased insurance costs, driver retention issues, and reputational risk. The ability to identify interpretable, high-impact crash risk factors provides actionable insights to transportation managers and safety analysts to design more targeted interventions. SHAP-RFE transforms feature selection from a technical preprocessing step into a diagnostic tool that supports real-world decision-making and operational resilience.

5.3 Future Work

Future work will focus on leveraging the selected SHAP-RFE features to build a fully optimized injury severity prediction model. This includes hyperparameter tuning, model comparisons (e.g., Random Forest, SVM), and evaluation across additional FARS datasets from different years. Broader extensions may involve applying the SHAP-RFE framework to other vehicle types, integrating SHAP-based insights into risk monitoring dashboards, and generating actionable guidance for fleet operators or transportation policymakers based on model outputs.

References

[1] American Trucking Associations. 2019. *Freight Transportation Forecast to 2029.* https://www.trucking.org/article/ATA-U.S.-Freight-Transportation-Forecast-to-2029. Accessed: 2025-05-24.

[2] National Safety Council. 2023. *Injury Facts: Large Trucks.* https://injuryfacts.nsc.org/motor-vehicle/road-users/large-trucks/. Accessed: 2025-05-24.

[3] Research and Markets. 2023. *Global Freight Trucking Market Report: 2023–2030 Industry Trends, Opportunities, and Forecasts.* https://www.researchandmarkets.com/report/trucking. Accessed: 2025-05-24.

[4] Jing Huang, Yang Peng, and Lin Hu. 2024. A multilayer stacking method based on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state. *Expert Systems with Applications 238* (2024), 121729. https://doi.org/10.1016/j.eswa.2023.121729. Accessed: 2025-05-24.

[5] Zhen Ma, Guoqing Mei, and Salvatore Cuomo. 2021. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis Prevention 160* (2021), 106322. https://doi.org/10.1016/j.aap.2021.106322.

[6] M. Osman, K. H. M. A. Rahman, and M. A. Abdel-Aty. 2016. Analysis of injury severity of large-truck crashes in work zones. *Accident Analysis Prevention 97* (2016), 261–273. https://doi.org/10.1016/j.aap.2016.09.012.

[7] A. Behnood and N. S. S. Al-Bdairi. 2020. Determinants of injury severities in large truck crashes: A weekly instability analysis. *Safety Science 131* (2020), 104911. https://doi.org/10.1016/j.ssci.2020.104911.

[8] Eric R. Teoh, Daniel L. Carter, Sarah Smith, and Anne T. McCartt. 2017. Crash risk factors for interstate large trucks in North Carolina. *Journal of Safety Research 62* (2017), 13–21. https://doi.org/10.1016/j.jsr.2017.05.002.

[9] Zijian Zheng, Pan Lu, and Brenda Lantz. 2018. Commercial truck crash injury severity analysis using gradient boosting data mining model. *Journal of Safety Research 65* (2018), 115–124. https://doi.org/10.1016/j.jsr.2018.03.002.

[10] Chao Yang, Mingyang Chen, and Quan Yuan. 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis Prevention 158* (2021), 106153. https://doi.org/10.1016/j.aap.2021.106153.

[11] Ismail Abdulrashid, Reza Zanjirani Farahani, Shamkhal Mammadov, Mohamed Khalafalla, and Wen-Chyuan Chiang. 2024. Explainable artificial intelligence in transport logistics: Risk analysis for road accidents. *Transportation Research Part E: Logistics and Transportation Review 186* (2024), 103563. https://doi.org/10.1016/j.tre.2024.103563.

[12] Jasmine Pahukula, Salvador Hernandez, and Avinash Unnikrishnan. 2015. A time of day analysis of crashes involving large trucks in urban areas. *Accident Analysis Prevention 75* (2015), 155–163. https://doi.org/10.1016/j.aap.2014.11.021.

[13] Nabeel Saleem Saad Al-Bdairi and Salvador Hernandez. 2017. An empirical analysis of run-off-road injury severity crashes involving large trucks. *Accident Analysis Prevention 102* (2017), 93–100. https://doi.org/10.1016/j.aap.2017.02.024.

[14] Shuo Zhang and Mohamed Abdel-Aty. 2022. Real-time crash potential prediction on freeways using connected vehicle data. *Analytic Methods in Accident Research 36* (2022), 100239. https://doi.org/10.1016/j.amar.2022.100239.

[15] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30* (2017). https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[16] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8* (2007), 25. https://doi.org/10.1186/1471-2105-8-25.

[17] Wei Zou, Xiaokun Wang, and Dapeng Zhang. 2017. Truck crash severity in New York City: An investigation of the spatial and the time of day effects. *Accident Analysis Prevention 99* (2017), 249–261. https://doi.org/10.1016/j.aap.2016.11.024.

[18] Mouyid Islam, Parisa Hosseini, and Mohammad Jalayer. 2022. An analysis of single-vehicle truck crashes on rural curved segments accounting for unobserved heterogeneity. *Journal of Safety Research 80* (2022), 148–159. https://doi.org/10.1016/j.jsr.2021.11.011.

[19] Muhammad Tahmidul Haq, Milan Zlatkovic, and Khaled Ksaibati. 2020. Investigating occupant injury severity of truck-involved crashes based on vehicle types on a mountainous freeway: A hierarchical Bayesian random intercept approach. *Accident Analysis Prevention 144* (2020), 105654. https://doi.org/10.1016/j.aap.2020.105654.

[20] Bhaven Naik, Li-Wei Tung, Shanshan Zhao, and Aemal J. Khattak. 2016. Weather impacts on single-vehicle truck crash injury severity. *Journal of Safety Research 58* (2016), 57–65. https://doi.org/10.1016/j.jsr.2016.06.005.