Crash Severity among At-Fault and Non-Fault Drivers: a Comparative Study

Febri Kurniawan [a,1*]

[a]School of Traffic and Transportation Engineering, Central South University, Changsha, China
[1*]febrikurniawan273@gmail.com

[*]corresponding e-mail

ABSTRACT

Traffic crashes remain a global public safety concern, with varying levels of injury severity and vehicle damage influenced by driver fault status. This study aims to compare crash severity between at-fault and non-fault drivers and identify the key contributing factors. Using Japanese traffic crash data from 2019 to 2022, the study combined the analysis of Generalized Estimating Equations (GEE) for interpretable statistical relationships and Random Forest (RF) for high-accuracy predictive model. Class imbalance was addressed through custom resampling and extensive feature engineering was applied. The results show that at-fault drivers are more likely to suffer severe injuries, especially due to stop sign violations, dangerous impact locations, and unsafe road conditions, while non-fault drivers tend to experience greater vehicle damage, particularly in rear-end and side-impact crashes. RF achieved 71% accuracy in predicting injury severity and 65% for vehicle damage classification, outperforming traditional models, while GEE confirmed the influence of traffic compliance, driver age, and crash context. These findings highlight the need for stricter enforcement of traffic regulations, intersection re-designs, and age-specific safety programs to reduce crash severity and improve road safety outcomes.

Keywords: traffic crash severity, at-fault drivers, non-fault drivers, generalized estimating equations, random forest

A. Introduction

Traffic safety remains a major global concern, particularly in urban areas where increasing vehicle density and human decision-making contribute to rising crash rates. Road crashes result in injuries and fatalities, and significant economic and public health impacts (Adegbite et al., 2019). While numerous studies have analyzed crash causation, the role of driver fault status, whether at-fault or non-fault, in determining crash severity remains understudied (Kitali, 2023; Taylor et al., 2018). Understanding how different driver categories experience crashes is essential for developing data-driven safety policies.

Japan offers a valuable case study due to its well-documented traffic system and comprehensive crash reporting (Shoots-Reinhard & Shihab, 2020). Unlike many Western datasets that often omit explicit fault classification, Japan's detailed records allow for an in-depth analysis of how driver responsibility correlates with crash severity. Additionally, demographic factors such as driver age and experience significantly influence crash risk, with younger drivers being disproportionately involved in severe crashes (Nguyen & Vu, 2023). Understanding these disparities is crucial for developing effective safety interventions.

This study aims to compare crash severity between at-fault and non-fault drivers to identify which group is at greater risk of severe outcomes. Additionally, it analyzes key factors influencing crash severity, such as traffic control compliance, impact location, and road conditions, to determine the most significant contributors to crash severity. Furthermore, this research examines the role of driver demographics, particularly age and experience, in shaping crash risk and severity. These findings highlight variations in crash outcomes among different driver groups.

To investigate these aspects, this study employs Generalized Estimating Equations (GEE) to examine the influence of correlated factors on crash severity and Random Forest (RF) to identify the most significant predictors of severe crashes. These methods allow for a robust analysis of both statistical relationships and complex non-linear interactions in the dataset (Adanu et al., 2021; Meng et al., 2020).

The findings of this study have critical implications for road safety strategies and policymaking. By identifying high-risk driver groups and key crash determinants, transportation authorities can implement targeted interventions such as stricter law enforcement for at-fault drivers, improved intersection safety for non-fault drivers, and tailored safety programs for younger drivers. Additionally, these insights can inform global traffic safety efforts by providing a framework for reducing crash risks in urban environments worldwide (Rezapour et al., 2019).

Through this analysis, the study contributes to the existing literature by offering a structured comparison of crash risks based on driver fault status and providing actionable insights for policymakers. Addressing urban traffic safety requires a continuous evidence-based approach to ensure effective interventions that reduce crash severity and enhance road safety outcomes.

B. Methods

1. Data Selection and Pre-processing

This study utilized Japanese traffic crash reports from 2019 to 2022, which provided detailed records of at-fault and non-fault drivers, allowing for a structured comparison of crash severity. The dataset was cleaned and validated by removing erroneous, incomplete, and duplicate entries. Only car-to-car collision cases with complete information on injury severity and vehicle damage were retained to ensure the dataset remained representative (Katayama et al., 2019).

To enhance data integrity, numerical values were standardized, and categorical variables were transformed using One-Hot Encoding to prevent ordinal biases (Weerasekera et al., 2020). Additionally, Winsorization was applied to limit the influence of extreme values in key numerical variables such as speed limits, lane width, and driver age, ensuring that the dataset more accurately reflected real-world conditions (Wu, 2024).

2. Initial Data Preparation, Variable Selection, and Feature Engineering

A comprehensive methodology for pre-processing, feature selection, and feature engineering was applied to improve the accuracy of traffic crash predictions, with some stages such as; a. Feature Engineering and Outlier Handling; b. Feature Selection and Multicollinearity Handling; c. Addressing Data Imbalance and Kolmogorov-Smirnov Test; d. Principal Model Testing; and e. Evaluation Metrics

Overall, this custom resampling method efficiently addressed class imbalance while maintaining the integrity of the dataset. The balanced data ensures reliability for subsequent predictive modeling and minimizes biases in model training and evaluation.

a Generalized Estimating Equations (GEE)

Generalized Estimating Equations (GEE) is a statistical method designed to analyse correlated outcomes by incorporating a working correlation matrix into the estimation process (Kassim et al., 2014). In contrast, the unstructured structure allowed correlations to vary freely between pairs of observations

Table 1. Dependent Variables Before and After Resampling

Dependent Variables	Class	Before Resampling	After Resampling
Injury severity (at-fault)	Injury	35,359	60,000
	No injury	98,089	60,000
Injury severity (non-fault)	Injury	103,022	60,000
	No injury	30,426	60,000
Vehicle damage (at-fault)	Damage	41,608	60,000
	No damage	91,840	60,000
Vehicle damage (non-fault)	Damage	47,129	60,000
	No damage	86,319	60,000

within a group, providing greater flexibility in modeling complex dependency patterns (Thijssen et al., 2017).

GEE provides unbiased estimates of predictor effects, even when the correlation structure is misspecified, making it highly suitable for complex datasets with traffic crash dependencies (Lipsitz et al., 2017).

b) *Random Forest (RF)*

To complement the statistical analysis, Random Forest (RF) was incorporated as a predictive modeling technique. RF is known for its high accuracy and robustness in handling large datasets with complex, high-order interactions (Khalilabadi et al., 2024).

RF is particularly useful for mitigating overfitting and handling imbalanced classes, as it employs bootstrap samples and feature selection techniques to enhance model performance (Wan et al., 2023).

To ensure the robustness, reliability, and interpretability of the models, multiple evaluation metrics and validation techniques were employed, with some stages such us; 1) Classification Performance Metrics. The ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) quantified the model's ability to distinguish between different levels of crash severity, providing an aggregate measure of classification performance across various threshold settings (Mousavi & Ouenniche, 2018), 2) Model Fit and Complexity Analysis. and 3) Error Analysis and Overfitting Prevention

To assess error patterns and prevent overfitting, several validation techniques were employed. Confusion matrices were examined to identify misclassification trends and gain insights into areas where model performance could be improved (Pradhan & Bhattacharya, 2020). Additionally, Pseudo-R^2 values were calculated to measure how well the independent variables accounted for variations in injury severity and vehicle damage, offering a quantitative measure of model explanatory power (Brandstätter et al., 2015). Finally, 5-fold cross-validation was conducted to test the generalizability of the models, ensuring that performance remained consistent across different data subsets and reducing the likelihood of overfitting.

3. Results and Disccussion

This section presents the findings by focusing on two key aspects: injury severity differences between at-fault and non-fault drivers and vehicle damage severity analysis. The statistical models used, Generalized Estimating Equations (GEE) and Random Forest (RF), are integrated within these sections to provide both interpretative and predictive insights.

1. Injury Severity: At-Fault vs. Non-Fault Drivers

The severity of injuries are compared between at-fault and non-fault drivers to determine which group faces a higher risk of severe outcomes. The results from the GEE and RF models are summarized in Table 2.

Before resampling, the GEE model shows that at-fault drivers had a higher

Table 2. Injury Severity Analysis for At-Fault vs. Non-Fault Drivers

Predictor	GEE (Before Resampling, β, p-value)	GEE (After Resampling, β, p-value)	RF Feature Importance
Age (at-fault)	-0.189, p<0.001p < 0.001p<0.001	-0.143, p<0.001p < 0.001p<0.001	0.063
Stop Sign (at-fault) no regulations	0.882, p<0.001p < 0.001p<0.001	0.569, p<0.001p < 0.001p<0.001	0.126
Point of Impact (at-fault) rear-end	0.447, p<0.001p < 0.001p<0.001	0.300, p<0.001p < 0.001p<0.001	0.099
Traffic Signal out of order signal	-0.262, p<0.001p < 0.001p<0.001	0.248, p<0.001p < 0.001p<0.001	0.081

likelihood of sustaining severe injuries, with key predictors being stop sign violations, collision impact location, and road conditions. The analysis also suggests that older drivers were less prone to severe injuries ($\beta = -0.1899$, $p < 0.001$), likely due to more cautious driving behaviour.

After resampling, the pattern remains similar with reduced coefficient magnitudes, indicating that data imbalance had previously exaggerated some effects. The RF model confirms the importance of stop sign violations, point of impact, and traffic signals in predicting severe injury cases. The recall for severe injuries has improved from 36% before resampling to 73% after resampling, indicating a much better ability to classify high-risk cases.

These findings suggest that at-fault drivers generally face a higher risk of injury, but non-fault drivers are also at risk when involved in crashes due to hazardous road conditions and improper traffic controls.

2. Vehicle Damage Severity Analysis

The severity of vehicle damage is analyzed separately for at-fault and non-fault drivers, considering factors such as collision point, lane width, and road conditions. The GEE and RF results are presented in Table 3.

The GEE results before resampling show that the point of impact, lane width, and traffic signals significantly influenced vehicle damage. The age of the driver negatively has impacted the vehicle damage ($\beta = -0.0416$), suggesting that older drivers experienced less severe vehicle damage, possibly due to more defensive driving.

After resampling, the relationship between age and damage severity is flipped ($\beta = 0.0207$), indicating that older drivers may still experience considerable vehicle damage despite lower injury severity. The RF model confirms the importance of lane width and traffic signals, indicating that road infrastructure plays a crucial role in vehicle damage severity.

Overall, vehicle damage is more severe for non-fault drivers, possibly due to the nature of rear-end and side-impact crashes, where non-fault drivers have less control over the collision.

Table 3. Vehicle Damage Analysis for At-Fault vs. Non-Fault Drivers

Predictor	GEE (Before Resampling, β, p-value)	GEE (After Resampling, β, p-value)	RF Feature Importance
Age (at-fault)	-0.042, p<0.001 p < 0.001 p<0.001	0.021, p<0.001p < 0.001p<0.001	0.054
Lane Width	0.108, p<0.001 p < 0.001 p<0.001	0.061, p<0.001p < 0.001p<0.001	0.054
Point of Impact (at-fault) rear-end	-0.253, p<0.001 p < 0.001 p<0.001	-0.151, p<0.001 p < 0.001 p<0.001	0.055
Traffic Signal out of order signal	-0.135, p<0.001 p < 0.001 p<0.001	-0.182, p<0.001 p < 0.001 p<0.001	0.082

3. Summary of Findings

By restructuring the results around at-fault versus non-fault drivers and vehicle damage, the findings offer a clearer understanding of which driver group faces higher risks in traffic crashes. The analysis reveals that at-fault drivers are more likely to sustain severe injuries, particularly in cases involving stop sign violations, hazardous road conditions, and critical points of impact during collisions. In contrast, non-fault drivers often experience more substantial vehicle damage, which is frequently associated with rear-end and side-impact crashes. These types of collisions typically occur in situations where non-fault drivers have limited control over the outcome.

Traffic control factors, such as stop signs and signals, play a major role in both injury and damage severity. Older drivers generally experience lower injury severity but may still face considerable vehicle damage. These insights provide important implications for road safety measures, such as targeted traffic control interventions and stricter at-fault driver monitoring.

This study examines injured severity and vehicle damage severity by comparing at-fault and non-fault drivers, utilizing Generalized Estimating Equations (GEE) and Random Forest (RF). The results provide key insights into which driver group is at greater risk and which factors most significantly influence crash outcomes. The following sections discuss the comparative performance of these methods, the most significant determinants of crash severity, practical implications, study limitations, and recommendations for future research.

4. Comparative Performance of GEE and RF in Injury and Vehicle Damage Analysis

GEE and RF were both applied to analyze injury severity and vehicle damage severity for at-fault and non-fault drivers. GEE proved effective in handling correlated crash outcomes, while RF excelled in predictive accuracy by capturing complex variable interactions. For injury severity, RF achieved the highest accuracy (71% after resampling) and recall (73% for severe injuries), demonstrating its ability to identify high-risk cases. GEE, on the other hand, provided interpretable coefficients that highlighted significant risk factors, such as stop sign violations, point of impact, and age of the driver.

For vehicle damage severity, RF also outperformed statistical models, with 65% accuracy after resampling, improving the classification of severe vehicle damage cases. Lane width, impact location, and compliance with traffic signals were the most critical factors identified by both models. GEE reinforced these findings by providing statistically significant relationships, confirming that non-fault drivers often experience more severe vehicle damage due to the nature of crashes.

Although RF demonstrated superior accuracy, its lack of interpretability poses a challenge for direct policy applications. Conversely, GEE offered statistically validated associations but was less effective in modeling complex, non-linear relationships. A hybrid approach combining RF's predictive strengths with GEE's explanatory power could enhance both model accuracy and policy relevance.

5. Key Determinants of Injury and Vehicle Damage Severity

a) *Injury Severity and At-Fault vs. Non-Fault Drivers*

The results indicate that at-fault drivers face a higher risk of severe injuries, primarily due to stop sign violations, point of impact during collision, and road conditions. RF ranked stop sign violations as the most critical predictor, followed by impact location and driver age. GEE confirmed that older drivers were less likely to sustain severe injuries, suggesting that cautious driving behavior mitigates injury risks.

However, non-fault drivers are not exempt from injury risks, especially in cases where traffic control devices fail or visibility is obstructed. The impact location (particularly right-side collisions) is a significant determinant of injury severity across both models. These findings align with prior studies highlighting the role of intersection design and signal compliance in reducing injury severity (Dingus et al., 2016).

b) *Vehicle Damage and Crash Severity*

The severity of vehicle damage is notably higher for non-fault drivers, especially in rear-end and side-impact crashes. GEE identifies lane width and point of impact as significant contributors, while RF highlights traffic signal compliance and impact location as key factors. Interestingly, the age of the driver has a varying effect on vehicle damage severity. Before resampling, older drivers appeared to have less severe vehicle damage, but after resampling, this effect diminished. This suggests that while defensive driving reduces crash impact, certain collisions (e.g., high-speed impacts) may still lead to significant vehicle damage, even for older drivers.

These findings emphasize the importance of road infrastructure improvements, particularly at high-risk intersections, to minimize both injury severity and vehicle damage outcomes.

D. Conclusion

This study examined injury severity and vehicle damage severity among at-fault and non-fault drivers, leveraging both Generalized Estimating Equations (GEE) and Random Forest (RF) to analyze risk factors and predictive performance. The findings highlight significant differences between at-fault and non-fault drivers in crash outcomes and reinforce the importance of traffic compliance, intersection safety, and driver demographics in determining severity levels.

The results confirm that at-fault drivers are more likely to sustain severe injuries, primarily due to stop sign violations, hazardous impact locations, and unsafe road conditions. In contrast, non-fault drivers experience greater vehicle damage severity, particularly in side-impact and rear-end collisions, which are often unavoidable due to other drivers' negligence. These findings emphasize the need for stricter enforcement of stop sign violations and improved road infrastructure at high-risk intersections.

Methodologically, RF outperformed statistical models in predictive accuracy, achieving 71% accuracy for injury severity and 65% for vehicle damage after resampling. Its ability to capture nonlinear relationships and variable interactions makes it a valuable tool for crash prediction. However, its lack of interpretability limits direct policy applications. On the other hand, GEE provided statistically robust insights, confirming the significance of stop sign violations, impact location, and driver age, but was less flexible in handling complex interactions. A hybrid approach combining RF's predictive strengths with GEE's interpretability could enhance both analysis and policymaking.

The study's findings highlight the critical role of infrastructure maintenance, compliance with traffic control devices, and the need for interventions tailored to specific driver demographics. To enhance road safety, traffic management strategies should give priority to several key areas. First, stronger enforcement and consistent maintenance of traffic signals and stop signs are essential, particularly at high-risk intersections where violations frequently lead to severe crashes. Second, intersection redesigns should be considered to reduce the occurrence of side-impact and rear-end collisions, which tend to disproportionately impact non-fault drivers who have limited control over the crash dynamics. Third, safety programs should be adapted to different age groups. For example, defensive driving courses targeting younger drivers can help reduce risk-taking behaviors, while infrastructure improvements that enhance visibility and navigation can support the safety of senior drivers. Lastly, the integration of machine learning models into real-time crash risk prediction offers a valuable tool for authorities to anticipate high-risk situations and deploy targeted, proactive interventions more effectively.

This study provides valuable insights, but some limitations persist, including its focus on Japan, the exclusion of environmental and behavioral factors, and the interpretability challenges associated with machine learning models. Future research should incorporate real-time traffic data, expand variable selection, explore hybrid modeling frameworks, and perform cross-regional analyses to enhance generalizability.

The integration of statistical methods with machine learning enhances crash risk prediction, facilitates data-driven policymaking, and improves road safety outcomes.

E. References

Adanu, E. K., Agyemang, W., Islam, R., & Jones, S. (2021). A Comprehensive Analysis of Factors That Influence Interstate Highway Crash Severity in Alabama. *Journal of Transportation Safety & Security*, *14*(9), 1552–1576. https://doi.org/10.1080/19439962.2021.1949414

Adegbite, Q., Billah, K., Sharif, H. O., & Dessouky, S. (2019). Urban Intersections and Traffic Safety in the City of San Antonio. *Matec Web of Conferences*, *271*, 06003. https://doi.org/10.1051/matecconf/201927106003

Brandstätter, V., Giesinger, L., Job, V., & Frank, E. (2015). The Role of Deliberative Versus Implemental Mindsets in Time Prediction and Task Accomplishment. *Social Psychology*, *46*(2), 104–115. https://doi.org/10.1027/1864-9335/a000231

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Pérez, M. Á., Buchanan King, M., & Hankey, J. M. (2016). Driver Crash Risk Factors and Prevalence Evaluation Using Naturalistic Driving Data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Jalali Khalilabadi, P., Karimpour, A., & Wu, Y.-J. (2024). Severity analysis of red-light–running behavior at signalized intersections. *Journal of Transportation Safety & Security*, *16*(5), 482–506. https://doi.org/10.1080/19439962.2023.2232324

Kassim, S., Jemain, A. A., & Hasan, H. (2014). *Modeling the Correlation Between Binary Longitudinal Data*. https://doi.org/10.1063/1.4894347

Katayama, Y., Kitamura, T., Kiyohara, K., Sado, J., Hirose, T., Matsuyama, T., Izawa, J., Nakagawa, Y., & Shimazu, T. (2019). Prehospital Factors Associated With Death on Hospital Arrival After Traffic Crash in Japan: A National Observational Study. *BMJ Open*, *9*(1), e025350. https://doi.org/10.1136/bmjopen-2018-025350

Kitali, A. E. (2023). Safety Evaluation of the First Bus Rapid Transit System in Tanzania. *Transportation Research Record Journal of the Transportation Research Board*, *2678*(8), 155–170. https://doi.org/10.1177/03611981231214228

Lipsitz, S. R., Fitzmaurice, G. M., Sinha, D., Hevelone, N. D., Hu, J. C., & Nguyen, L. L. (2017). One-Step Generalized Estimating Equations With Large Cluster Sizes. *Journal of Computational and Graphical Statistics*, *26*(3), 734–737. https://doi.org/10.1080/10618600.2017.1321552

Meng, F., Xu, P., Song, C., Gao, K., Zhou, Z., & Yang, L. (2020). Influential Factors Associated With Consecutive Crash Severity: A Two-Level Logistic Modeling Approach. *International Journal of Environmental Research and*

Public Health, *17*(15), 5623. https://doi.org/10.3390/ijerph17155623

Mousavi, M. M., & Ouenniche, J. (2018). Multi-Criteria Ranking of Corporate Distress Prediction Models: Empirical Evaluation and Methodological Contributions. *Annals of Operations Research*, *271*(2), 853–886. https://doi.org/10.1007/s10479-018-2814-2

Nguyen, D. V. M., & Vu, A. T. (2023). Roadway Departure Crashes in Vietnam: Identifying the Sequence of Events, Causes, and Countermeasures by Applying the Crash Reconstruction Technique. *Iop Conference Series Materials Science and Engineering*, *1289*(1), 012050. https://doi.org/10.1088/1757-899x/1289/1/012050

Rezapour, M., Molan, A. M., & Ksaibati, K. (2019). Application of Multinomial Regression Model to Identify Parameters Impacting Traffic Barrier Crash Severity. *The Open Transportation Journal*, *13*(1), 57–64. https://doi.org/10.2174/1874447801913010057

Shoots, Reinhard, B., & Shihab, M. (2020). *Effect of Political Ideology, Stay-at-Home Rates, and Increased Risk Taking in Ohio Drivers During COVID-19 Shutdown*. https://doi.org/10.31234/osf.io/sq3bx

Taylor, S. G., Russo, B. J., & James, E. (2018). A Comparative Analysis of Factors Affecting the Frequency and Severity of Freight-Involved and Non-Freight Crashes on a Major Freight Corridor Freeway. *Transportation Research Record Journal of the Transportation Research Board*, *2672*(34), 49–62. https://doi.org/10.1177/0361198118776815

Thijssen, A., Creemers, A., Elst, W. V. d., Creemers, E., Vandormael, E., Dhont, N., & Ombelet, W. (2017). Predictive Value of Different Covariates Influencing Pregnancy Rate Following Intrauterine Insemination With Homologous Semen: A Prospective Cohort Study. *Reproductive Biomedicine Online*, *34*(5), 463–472. https://doi.org/10.1016/j.rbmo.2017.01.016

Wan, M., Wu, Q., Yan, L., Guo, J., Li, W., Lin, W., & Lu, S. (2023). Taxi drivers' traffic violations detection using random forest algorithm: A case study in China. *Traffic Injury Prevention*, *24*(4), 362–370. https://doi.org/10.1080/15389588.2023.2191286

Wang, S., Minku, L. L., & Yao, X. (2015). Resampling-Based Ensemble Methods for Online Class Imbalance Learning. *Ieee Transactions on Knowledge and Data Engineering*, *27*(5), 1356–1368. https://doi.org/10.1109/tkde.2014.2345380

Weerasekera, R., Sridharan, M., & Ranjitkar, P. (2020). Implications of Spatiotemporal Data Aggregation on Short-Term Traffic Prediction Using Machine Learning Algorithms. *Journal of Advanced Transportation*, *2020*, 1–21. https://doi.org/10.1155/2020/7057519

Wu, N. (2024). A Novel Hybrid Model Based on Rule Learning and Dilated Convolution for Vehicle Collision Prediction. *Ann Applied Sciences*. https://doi.org/10.55085/aas.2024.720