A STATISTICAL AND MACHINE LEARNING FRAMEWORK FOR ANALYZING ROAD TRAFFIC CRASH SEVERITY IN SRI LANKA

Dhanushi Bhagya[1*], Nalaka Jayantha[1], H.R. Pasindu[1]
[1]*Department of Civil Engineering, Faculty of Engineering, University of Moratuwa*

ABSTRACT

This study analyses 397,850 police-reported crashes in Sri Lanka from 2010 to 2020 using a multinomial logit model, validated through Random Forest and XGBoost, to identify key factors influencing crash severity across four outcome levels. Results reveal four findings not apparent from conventional crash statistics. First, motorcycle–vulnerable road user crashes (OR = 202, 10.24% of crashes) produced higher fatal odds than heavy vehicle–motorcycle interactions (OR = 143, 6.86%), despite the latter receiving greater policy attention. This suggests that motorcycle-related risk to vulnerable road users is considerably underappreciated in current safety programs. Second, single-vehicle motorcycle crashes (OR = 46.2, 9.05%) were three times more fatal per incident than motorcycle–passenger vehicle collisions (OR = 14.6), implicating loss of control and emergency response delays as under-recognized contributors. Third, passengers falling from buses yielded a fatal odds ratio of 2,337, the highest in the model, despite representing only 1.50% of crashes, indicating a concentrated fatality risk understated in national reporting. Fourth, dusk and dawn conditions (OR = 2.24, 17.27% of crashes) were found to be equally as dangerous as completely unlit nighttime roads (OR = 2.26), while fog and mist produced the highest lighting-related fatal odds ratio of 4.71. The findings indicate that targeted interventions addressing motorcycle–vulnerable road user interactions, bus passenger safety, and transitional lighting conditions offer the greatest potential for reducing fatal crash outcomes in Sri Lanka.

Keywords: Crash Severity, Multinomial Logit Model, Vehicle Interaction, Road Safety

1. INTRODUCTION

Road traffic crashes impose a substantial public health and economic burden globally, with the World Health Organization estimating 1.35 million fatalities annually. In Sri Lanka, crash fatalities and casualties have shown a persistent upward trend over the past two decades, reflecting growth in motorization that has outpaced corresponding advances in road safety infrastructure and enforcement. While considerable research has examined crash frequency, comparatively little attention has been directed at quantifying the factors that determine how severe a crash becomes once it has occurred. Understanding crash severity is critical because it informs the design of targeted interventions that can reduce the proportion of crashes resulting in grievous injury or death, rather than simply reducing crash counts.

Sri Lanka presents a unique road safety environment characterized by high motorcycle exposure, mixed-traffic conditions involving heavy vehicles and vulnerable road users sharing the same carriageway, and a public transport system in which passengers frequently travel on overcrowded buses. These conditions give rise to crash mechanisms and risk patterns that are not well represented in existing international literature. There is therefore a need for a data-driven, context-specific analysis that identifies and quantifies the factors most strongly associated with severe and fatal crash outcomes in Sri Lanka. This study addresses that need by developing a validated MNL model using national police-reported crash data spanning 2010 to 2020, with the aim of providing an interpretable, evidence-based framework to directly inform road safety policy and intervention priorities.

https://doi.org/10.31705/TRF.2026.4

2. LITERATURE REVIEW

The MNL model is one of the most widely used approaches for analyzing crash severity, as it accommodates multiple discrete outcome categories without imposing an ordinal structure on the dependent variable. (Chen & Fan, 2019) applied the MNL model to pedestrian-vehicle crashes in North Carolina, finding that vehicle type, pedestrian age, and lighting conditions significantly increased fatality probability, with heavy trucks and motorcycles associated with the highest fatal risks. Similarly, (Abrari Vajari et al., 2020) used the MNL model to evaluate motorcycle crash severity at Australian intersections, finding that multi-vehicle crashes, rural locations, and uncontrolled intersections increased fatal injury probability, while adverse weather conditions paradoxically reduced it by discouraging high-speed riding.

(Michalaki et al., 2015) applied a generalized ordered logit model to motorway crashes in England, finding that heavy goods vehicle involvement and driver fatigue substantially increased severity, particularly on hard shoulders. (Shiran et al., 2021) compared MNL models with machine learning approaches for crash severity prediction on California highways, finding that crash cause and number of vehicles involved were the most influential predictors across all models. (Alsaleh et al., 2025) demonstrated that pedestrian and cyclist involvement produced exceptionally high fatal odds ratios exceeding 350, underscoring the acute vulnerability of unprotected road users across different contexts. While these studies provide valuable methodological and contextual insights, none specifically addresses the South Asian Road environment, where the combination of heavy vehicle dominance, high motorcycle exposure, and overcrowded public transport creates distinct crash severity patterns. This study contributes to the literature by applying a validated MNL framework to Sri Lanka's national crash dataset, identifying crash mechanisms specific to this context that are largely absent from international research.

3. METHOD

National police-reported crash data for Sri Lanka from 2010 to 2020 were obtained, comprising 397,850 records after cleaning and processing. Crash severity was classified into four levels: damage-only, non-grievous injury, grievous injury, and fatal. Vehicle types were aggregated into six functional categories comprising passenger vehicles, motorcycles, three-wheelers, heavy vehicles, vulnerable road users, and other vehicles. Sixteen vehicle interaction categories were subsequently derived from these groups. Collision types, lighting conditions, weather conditions, traffic control, urban-rural setting, and road geometry variables were similarly grouped into analytically meaningful categories. Records with missing values in variables used for model estimation were excluded from the final modelling dataset. The dataset was divided into a 70% training set (278,495 records) and a 30% testing set (119,355 records) to enable out-of-sample validation. Four groups of explanatory variables were included in the MNL model: vehicle interaction type, collision mechanism, environmental conditions, and road geometry. Vehicle interaction types included heavy vehicle versus vulnerable road user, heavy vehicle versus motorcycle, motorcycle versus vulnerable road user, motorcycle versus passenger vehicle, and single-vehicle crashes, with passenger vehicle versus passenger vehicle crashes serving as the reference category. Collision types included passengers falling from buses, head-on collisions, cycle and motorcycle collisions, and collisions with fixed objects, with angle collisions as the reference. Environmental variables covered weather conditions, lighting conditions, and road setting (urban versus rural), while road geometry variables included straight roads, curves, T-junctions, cross-junctions, roundabouts, and rail crossings. The MNL model was estimated using maximum likelihood, with odds ratios calculated for each severity level relative to the damage-only reference outcome. Because fatal crashes represented only 7.03% of the dataset, a balanced MNL model was subsequently developed to address class imbalance.

Grievous and fatal crash observations within the training dataset were randomly oversampled with replacement until their frequencies approximated those of the majority injury classes. The balanced dataset was then used to estimate a second MNL model and evaluate the sensitivity of severity predictions to class imbalance. For cross-model validation, Random Forest and XGBoost models were trained using the same explanatory variables. Model performance was assessed using accuracy and fatal-crash recall, allowing comparison between statistical and machine-learning approaches and providing additional verification of the robustness of the identified crash severity factors.

4. LIMITATIONS

Several limitations should be considered when interpreting the findings. First, the analysis relies on police-reported crash records and may therefore be affected by underreporting, particularly for less severe crashes. Second, reporting practices may vary across regions and years, introducing potential reporting bias. Third, the database does not contain detailed behavioral variables such as actual travel speed, seatbelt use, helmet use, distraction, or fatigue, which have been identified as important severity determinants in previous studies. Finally, the study is based on observational data and therefore identifies statistical associations rather than direct causal relationships. Despite these limitations, the large national dataset and consistency across statistical and machine learning models provide confidence in the robustness of the principal findings.

5. RESULTS AND DISCUSSION

Vehicle interaction type was the dominant predictor of crash severity, confirmed consistently across the MNL, Random Forest, and XGBoost models. Heavy vehicle–vulnerable road user crashes produced the highest fatal OR at 611, yet account for only 3.51% of all crashes. More critically, motorcycle–vulnerable road user crashes, nearly three times more frequent at 10.24% of crashes, produced a fatal OR of 202, higher than heavy vehicle–motorcycle interactions (OR = 143). A vulnerable road user struck by a motorcycle therefore faces greater fatal risk per crash than a motorcyclist struck by a heavy vehicle, an interaction that current road safety programs largely overlook.

Single-vehicle motorcycle crashes (OR = 46.2) were three times more fatal per incident than motorcycle–passenger vehicle collisions (OR = 14.6), pointing to loss of control and emergency response delays rather than collision impact as the primary contributors. Passengers falling from buses produced the highest fatal OR in the model at 2,337, despite representing only 1.50% of crashes, a catastrophic but consistently understated risk specific to South Asian public transport conditions. Dusk and dawn conditions (OR = 2.24) carried equivalent fatal risk to completely unlit nighttime roads (OR = 2.26), while fog and mist produced the highest lighting-related OR of 4.71. Road geometry variables showed odds ratios below unity, indicating a marginal severity-reducing effect, likely through speed moderation. The balanced MNL model improved fatal crash recall from two percent to 54%, and fatal crash odds rose to 2.3 times the 2010 baseline by 2020, with motorcycle fatality odds increasing most steeply.

5.1 Model Validation

To assess the robustness of the findings, the multinomial logit (MNL) model was validated against Random Forest and XGBoost machine learning models. Because crash severity data are highly imbalanced, overall accuracy alone may not adequately reflect a model's ability to identify fatal crashes. Therefore, both overall accuracy and fatal-crash recall were examined.

Table 1. Comparative performance of statistical and machine-learning models

Model	Accuracy	Fatal Recall	Key Insight
MNL	0.60	0.02	High overall accuracy but poor fatal-crash detection
Balanced MNL	0.52	0.54	Significant improvement in identifying fatal crashes
Random Forest	0.52	0.48	Comparable performance with improved fatal detection
XGBoost	0.60	0.52	Confirms findings while maintaining high predictive accuracy

6. CONCLUSION

This study analysed 397,850 police-reported crashes in Sri Lanka from 2010 to 2020 to identify and quantify the key determinants of crash severity. The findings demonstrate that crash severity is governed primarily by vehicle mass disparity in interactions and collision mechanisms, rather than road geometry or weather conditions. Four findings carry direct policy implications. First, motorcycle–vulnerable road user crashes (OR = 202, 10.24% of crashes) produce a higher fatal risk per crash than heavy vehicle–motorcycle interactions (OR = 143), indicating that protective interventions must extend beyond heavy vehicle separation to address the motorcycle-dominated mixed-traffic environment. Second, single-vehicle motorcycle crashes are three times more fatal per incident than motorcycle–vehicle collisions, pointing to road surface quality and emergency response as under-addressed risk factors. Third, Passengers falling from buses produced the highest fatal odds ratio in the model (OR = 2,337), despite representing only 1.50% of all crashes. This finding indicates an extremely elevated fatality risk associated with this crash mechanism. However, the magnitude of the estimated odds ratio should be interpreted cautiously because rare crash categories can generate unstable coefficient estimates and sparse-data effects in multinomial logit models. Consequently, the result is best interpreted as evidence of a highly hazardous crash mechanism rather than as a precise estimate of effect magnitude. Nevertheless, the finding highlights a concentrated fatality risk associated with overcrowded public transport operations that remain largely overlooked in conventional road safety assessments. Fourth, dusk and dawn conditions carry a fatal risk statistically equivalent to unlit nighttime roads, indicating that lighting interventions should target transitional periods alongside rural unlit environments. These findings provide a quantitative evidence base for targeted road safety interventions in Sri Lanka, with the greatest potential for fatality reduction lying in motorcycle–vulnerable road user separation measures, bus passenger safety enforcement, and expanded lighting strategies that account for the full spectrum of low-visibility conditions.

7. REFERENCES

1. Abrari Vajari, M., Aghabayk, K., Sadeghian, M., & Shiwakoti, N. (2020). A multinomial logit model of motorcycle crash severity at Australian intersections. *Journal of Safety Research*, *73*, 17–24. https://doi.org/10.1016/j.jsr.2020.02.008
2. Alsaleh, R., Walia, K., Moshiri, G., & Alsaleh, Y. T. (2025). Traffic Collision Severity Modeling Using Multi-Level Multinomial Logistic Regression Model. *Applied Sciences (Switzerland)*, *15*(2). https://doi.org/10.3390/app15020838
3. Chen, Z., & Fan, W. (David). (2019). A multinomial logit model of pedestrian-vehicle crash severity in North Carolina. *International Journal of Transportation Science and Technology*, *8*(1), 43–52. https://doi.org/10.1016/j.ijtst.2018.10.001