EVALUATION OF CONTRIBUTING FACTORS FOR AT-FAULT MOTORCYCLE RIDER CRASH SEVERITY IN SRI LANKA USING REGRESSION AND MACHINE LEARNING TECHNIQUES

Sachini Shashiprabha Madushani[1*], H.R. Pasindu[1]
[1]*University of Moratuwa*

ABSTRACT

This study aims to investigate the contributions of roadway, environmental, vehicle, collision, and driver-related factors to the severity of at-fault motorcycle crashes using both Logistic Regression (LR) and machine learning (ML) approaches. Approximately 200,000 motorcycle-related crash records from 2008 to 2020, obtained from the Sri Lanka Accident Data Management System (SLADMS), were analysed. The study classified crashes into two categories: high-severity crashes, including fatal and grievous crashes, and low-severity crashes, including non-grievous and property damage only crashes. Five categories of contributing factors were considered: driver-related, roadway-related, vehicle-related, environmental, and collision-related factors. Six ML models, namely Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGB), K-Nearest Neighbour (KNN), Support Vector Classifier (SVC), and Artificial Neural Network (ANN) were compared with Logistic Regression using an 80/20 train-test split. Shapley Additive Explanations (SHAP) were employed to interpret the ML models. Among the evaluated models, the RF model demonstrated the best performance in predicting high-severity crashes, achieving a recall of 0.87, a precision of 0.60, and an F1-score of 0.71. SHAP analysis indicated that collision type (SHAP value = 0.085) was the most influential factor contributing to high-severity crashes, followed by the collided element (SHAP value = 0.068) and human pre-crash factors (SHAP value = 0.045). The findings suggest that explainable ML methods can be effectively used to assess crash severity and to facilitate more informed road safety decision-making.

Keywords: crash severity, motorcycle, logistic regression, machine learning, SHAP analysis

1. INTRODUCTION

Motorcycles account for about 58% of the total registered vehicle fleet in Sri Lanka. (Department of Motor Traffic, 2026). Over the past few decades, the model share of motorcycles has increased substantially, leading to a significant increase in fatal motorcycle-related crashes. As motorcyclists constitute a vulnerable road user group, it is essential to investigate the root causes of these crashes to enhance the road safety management practises in Sri Lanka. Therefore, this study aims to analyse the patterns of crash severity among at-fault motorcycle riders in Sri Lanka, to develop and compare statistical and machine learning (ML) models for predicting crash severity, and to identify high-risk groups and the key factors contributing to the observed severity patterns. The results of the study can be utilised to improve road safety by providing accurate explanations for at-fault motorcycle crash severity. Finally, the findings would be useful to identify effective measures to improve the motorcycle riders' safety in Sri Lanka.

2. LITERATURE REVIEW

In the Sri Lankan context, studies have been conducted to evaluate the risk factors in crashes involving motorcycles. Shajith et al. (2019) showed that high-severity crash outcomes were more likely to occur when motorcycles were involved in head-on collisions and collisions with heavy vehicles or hard objects, using a stepwise binary Logistic Regression (LR) approach. Moreover, Amarasingha (2021) observed that riding on rural roadways, riding during weekdays,

https://doi.org/10.31705/TRF.2026.12

and riding newer motorcycles were among the most prominent factors associated with motorcycle crashes, based on an analysis of the odds ratios associated with environmental, roadway, rider, and vehicle characteristics. In the global context, novel techniques, including ML models, have been widely employed for similar analyses. Rezapour et al. (2021) and Wahab & Jiang (2019) used ML models and observed that the Random Forest (RF) model surpasses the other models, Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), and Support Vector Classifier (SVC) in predicting motorcycle crash injury severity. Therefore, this study was undertaken to address the existing research gap within the Sri Lankan context.

3. METHOD

3.1. Data Collection

Data for this study were obtained from the Sri Lanka Accident Data Management System (SLADMS) which contains information on police-reported road crashes across all severity levels in Sri Lanka between 2008 and 2020. The SLADMS crash data are organised into three separate files: (i) accident details, (ii) casualty details, and (iii) vehicle details. Based on the objectives of this study and the relevance of the information contained within each file, a customised database was developed. Relevant attributes from the vehicle details file were extracted and systematically appended to the corresponding accident records in the accident details file using the unique Crash Identification Number and the year of occurrence. This data integration process resulted in the creation of a consolidated database of crashes that captured crash-level, vehicle, and human characteristics associated with crash occurrence and severity. The resulting merged database formed the basis for subsequent data screening, preprocessing, and statistical analysis conducted in this study.

3.2. Data Pre-processing

After developing the integrated crash database, a systematic preprocessing procedure was conducted to ensure data consistency, reliability, and relevance to the study objectives. Crashes involving motorcycles as the at-fault party were identified using the human pre-crash factor and were subsequently extracted from the database. Data cleaning was then performed to remove records with missing, incomplete, incorrect, or inconsistent information, resulting in a final dataset suitable for analysis. Finally, continuous variables were converted into categorical variables to improve interpretability and facilitate more meaningful comparisons between groups, consistent with approaches adopted in previous studies (Shajith et al., 2019).

3.3. Variable Selection

Accident severity was selected as the dependent variable to investigate the factors influencing crash severity. In the database, crash severity was classified into four categories: fatal, grievous, no-grievous and property damage only. Crashes categorised as property damage only were excluded from the analysis, as these incidents are often underreported (Amarasingha, 2021). Therefore, the dataset was limited to crashes that involved human injuries, which are generally associated with more reliable reporting and higher data accuracy. For analytical purposes, crash severity was classified into two classes and binary coded: high severity crashes (coded as 1) – comprising fatal and grievous crashes, and low severity crashes (coded as 0) – comprising non-grievous crashes. The selection of this classification helped achieve class balance within the dataset. A set of independent variables representing different dimensions of crash characteristics was selected from the dataset, and the attributes corresponding to these independent variables were encoded using integer values. The independent variables considered in the study, along with their respective attributes and encoded values, are presented in Table 1.

Table 1. Variables and attributes used in the study

Factor	Variable	Values and Attributes
Driver	Gender	0 = Male, 1 = Female
	Age	0 = <18 years (yrs), 1 = 18-25 yrs, 2 = 26-40 yrs, 3 = 41-60 yrs, 4 = >60 yrs
	Validity of license	0 = International license/ Valid license for the vehicle, 1 = Probation license/ Learner permit, 2 = Without valid license for the vehicle
	Driving Experience	0 = 0-5 yrs, 1 = 6-10 yrs, 2 = 11-20 yrs, 3 = >20 yrs
Vehicle	Age	0 = 0-5 yrs, 1 = 6-10 yrs, 2 = >10 yrs
Roadway	Region	0 = Urban, 1 = Rural
	Location type	0 = Mid-block, 1 = Access point, 2 = Junction, 3 = Rail Crossing
Environmental	Road surface condition	0 = Dry, 1 = Wet, 2 = Slippery surface (mud, oil, garbage, leaves), 3 = Flooded with water
	Weather condition	0 = Clear, 1 = Cloudy/Foggy, 2 = Rainy
	Lighting condition	0 = Daylight, 1 = Night with good street lighting, 2 = Dusk, dawn, 3 = Night with no or improper street lighting
Collision	Collision Type	01 = Head-on collisions, 02 = Collision with pedestrians/ Collision with animals/ Collision with cyclists, 03 = Rear-end collisions, 04 = Sideswipe collisions/ Side-impact collisions/ T-bone accidents (Opposite directions)/ T-bone accidents (same direction), 05 = Collision with stopped vehicle/ Single-vehicle collisions/ Collision with hard objects
	Collided Element	01 = Bus/ Lorry/ Land Vehicle, 02 = Car/ Dual purpose vehicle, 03 = Motorcycle/ Three-wheeler, 04 = Cycle/ Animal/ Pedestrian
	Human Factor	01 = Speeding, 02 = Aggressive/negligent driving, 03 = Error of judgement, 04 = Influenced by alcohol/drugs, 05 = Fatigue/fall asleep, 06 = Distracted/ inattentiveness, 07 = Poor eyesight, 08 = Sudden illness, 09 = Blinded by another vehicle/sun

3.4. Data Analysis

A LR model and six ML models were used for the data analysis. The ML models considered in this study were DT, RF, KNN, SVC, Extreme Gradient Boost (XGB), and Artificial Neural Network (ANN). For this analysis, 80% of the crash records were allocated to model training, while the remaining 20% were used for model validation. The dataset was not stratified by severity class during the train/test split because the classes were sufficiently balanced. Model development began with an empty model, and features were added sequentially using the forward feature selection technique to identify the most relevant variables. Hyperparameter tuning was subsequently performed to optimise model performance.

4. RESULTS AND DISCUSSION

The performance evaluation parameters for the LR and ML models are presented in Table 2. According to the results, the RF model demonstrated the best overall performance for crash severity prediction among the LR and ML models evaluated in this study. Although the SVC model achieved perfect recall (1.00) for the high-severity class, its relatively low precision (0.56) indicates a high rate of false positives, meaning that while it identified all actual severe crashes, it also incorrectly classified many non-severe crashes as severe, limiting its practical utility.

Subsequently, Shapley additive Explanations (SHAP) were used to explain the RF model, and the feature importance associated with the prediction of high-severity crashes (Class 1) is shown in Figure 1. According to the SHAP values, collision factor emerged as the most influential feature in predicting Class 1 outcomes, while collided element and human pre-crash factor were identified as the second and third most influential features, respectively.

Table 2. Performance evaluation parameters for the LR and ML models

Model	All classes			Class 1 (Fatal & Grievous)		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Logistic Regression (LR)	0.58	0.57	0.53	0.86	0.58	0.69
Decision Tree (DT)	0.59	0.58	0.56	0.82	0.60	0.69
Random Forest (RF)	0.60	0.60	0.55	0.87	0.60	0.71
Extreme Gradient Boost (XGBoost)	0.59	0.59	0.56	0.84	0.60	0.70
K-Nearest Neighbour (KNN)	0.57	0.56	0.56	0.67	0.60	0.63
Support Vector Classifier (SVC)	0.56	0.31	0.40	1.00	0.56	0.72
Artificial Neural Network (ANN)	0.59	0.56	0.54	0.54	0.60	0.70

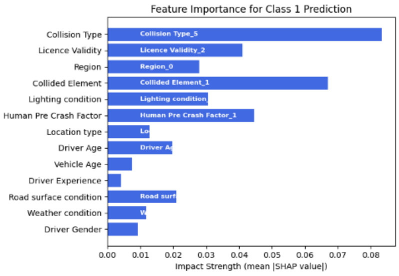

Figure 1. SHAP average feature importance for RF model

5. CONCLUSION

This study evaluated the effectiveness of explainable ML approaches for predicting traffic crash severity using roadway, environmental, vehicle, collision, and driver-related factors. Six ML models, namely RF, DT, SV, XGB, ANN and KNN, were compared with LR model using two crash severity categories: high severity (fatal and grievous) and low severity (non-grievous and property-damage-only). The RF model achieved the best predictive performance based on the evaluation metrics of recall, precision, and F1-score. SHAP-based interpretation revealed that collision type was the most influential factor affecting crash severity, followed by collided element and human pre-crash factors. The findings demonstrated that RF with SHAP provides superior predictive capability and enhanced interpretability for identifying high-severity crashes, thereby offering valuable insights for highway engineers and policymakers seeking to improve road safety decision-making.

6. REFERENCES

1. Amarasingha, N. (2021). Risk Factors of Crashes Involving Motorcycles in Sri Lanka. *Journal of South Asian Logistics and Transport*, *1*(2), 1–18. https://doi.org/10.4038/jsalt.v1i2.31
2. Department of Motor Traffic. (2026). *Vehicle Population 2014-2025*. Department of Motor Traffic. https://dmt.gov.lk/images/2026/Vehicle_Population_2014-2025.pdf
3. Rezapour, M., Farid, A., Nazneen, S., & Ksaibati, K. (2021). Using machine leaning techniques for evaluation of motorcycle injury severity. *IATSS Research*, *45*(3), 277–285. https://doi.org/10.1016/j.iatssr.2020.07.004
4. Shajith, S. L. A., Pasindu, H. R., & Ranawaka, R. K. T. K. (2019). Evaluating the Risk Factors in Fatal Accidents involving Motorcycle – Case Study on Motorcycle Accidents in Sri Lanka. *Engineer: Journal of the Institution of Engineers, Sri Lanka*, *52*(3), 33. https://doi.org/10.4038/engineer.v52i3.7363
5. Wahab, L., & Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLOS ONE*, *14*(4), e0214966. https://doi.org/10.1371/journal.pone.021496