



Prediction and Investigation of the Injury Severity of Drivers Involved in Speeding-Related Crashes Using Machine Learning Models

Neero Gumsar SORUM¹, Martina Gumsar SORUM²

Original Scientific Paper
Submitted: 18 Dec 2024
Accepted: 11 Mar 2025

¹ Corresponding author, neerogsorum@gmail.com, Department of Civil Engineering, North Eastern Regional Institute of Science and Technology, Nirjuli, India
² tinagsorum@gmail.com, Department of Civil Engineering, North Eastern Regional Institute of Science and Technology, Nirjuli, India



This work is licensed
under a Creative
Commons Attribution 4.0
International Licence.

Publisher:
Faculty of Transport
and Traffic Sciences,
University of Zagreb

ABSTRACT

Speeding is the major reason for road traffic crashes and deaths in India. The other driver's faults include driving under the influence, using mobile phones while driving and driving on the wrong side of the road. Therefore, this study attempts to predict and investigate the driver injury severity (DIS) in speeding-related crashes. A total of 793 police-reported single-vehicle and two-vehicle crash data from Imphal City, India, collected between 2011–2020, were analysed and modelled. For DIS prediction, eleven supervised machine learning (ML) models were implemented using 5-fold and 10-fold cross-validation (FCVs) and trained at train ratio (TR) values of 0.5, 0.6, 0.7 and 0.8 in each FCV. The top ML model for the DIS prediction was selected based on the best combination of recall, accuracy, F1 score, area under the curve (AUC) and precision metrics. Feature importance analysis (FIA) was conducted to determine the impactful factors in DIS prediction. The gradient boosting tree (GBT), stochastic gradient descent, decision tree and lasso-LARS models were identified as the top-performing ML models for the DIS prediction at TR = 0.5, 0.6, 0.7 and 0.8, respectively, in 5-FCV. The light GBM (TR = 0.5 and 0.7), GBT (TR = 0.6) and lasso-LARS (TR = 0.8) were the best-performing ML models in 10-FCV. The FIA results indicated that vehicle type (two-wheeler), nature of crash (head-on collision) and time of crash (12 PM–6 PM and 6 AM–12 PM) variables were the most impactful variables on the DIS prediction in Imphal speeding-related crashes. These ML models can be employed in hilly areas for the accurate prediction of DIS. The study results can help transportation planners in designing road safety measures and strategies to lessen DIS in speeding-related crashes.

KEYWORDS

speeding; driver injury severity; machine learning; Dataiku; gradient boosting tree; lasso-LARS.

1. INTRODUCTION

Road traffic accidents (RTAs) continue to pose a serious public health challenge in India, contributing to one out of every ten road traffic fatalities worldwide [1]. The country has shared 11% of worldwide RTA fatalities, imposing an economic burden approximately equal to 3% of the country's gross domestic product [2]. Though India initiated many road safety programs in the past decades, approximately 0.15 million people die on Indian roads annually, which translates, on average, into 1,130 crashes and fatalities every day or 47 crashes and 18 deaths per hour [2]. This figure will further increase unless effective road safety countermeasures are taken by the Government of India because the country's population, road construction and number of vehicles are increasing day by day. Moreover, over-speeding was reported to be the major reason for road crashes and deaths in India [2]. In the year 2021, around 74.4% of crashes and 72.2% of deaths were recorded due to speeding. The other major factors that cause RTAs and fatalities due to the driver's fault include driving under the influence of alcohol/drugs, mobile phone usage, driving on the wrong side and

jumping a red light [2]. Based on this information, examination of driver injury severity (DIS) in speeding-related crashes in India is of considerable importance.

Methodologically, in recent years, many researchers in the field of crash severity prediction and analysis have favoured machine learning (ML) technologies over classic statistical methods (logistic regression, ordered logit and ordered probit models) because the statistical models typically rely on predefined assumptions (e.g. linearity, normality, homoscedasticity) and often focus on capturing simpler relationships between variables. Moreover, these models struggle with large datasets or high-dimensional data. In addition, these models fail to explicitly account for unobserved heterogeneity (factors that influence outcomes but are not directly measured in the data) [3–5]. This leads to biased results and inaccurate conclusions. As a result, ML methods are often preferred in situations where predictive accuracy, no pre-defined assumptions, flexibility and the ability to model complex interactions are crucial [6]. To overcome the black box characteristic of the ML methods, several recent studies have employed the Shapley Additive exPlanations (SHAP) approach in crash severity prediction [7, 8]. SHAP is an application of cooperative game theory, specifically its solution concept known as Shapley values. SHAP does this by giving each feature in the input data point its score based on how much that feature contributes to the model's final prediction compared to the mean or average prediction made using all subsets of that feature across the dataset. This allows for a detailed, consistent and fair explanation of how each feature impacts the final prediction [9].

Therefore, the ML method is used in the present study to predict the DIS in speeding-related crashes and simultaneously to explore the impacts of influential factors on DIS prediction using feature importance analysis (FIA). The structure of the article is as follows: Section 2 provides a review of pertinent literature. The study data and methodology are elaborated in Section 3. Results and discussion of DIS prediction and FIA are presented in Section 4. Section 5 presents the summary and conclusions of the study, along with limitations and future studies.

2. LITERATURE REVIEW

Various ML methods have been used for the analysis and prediction of crash severity and DIS in several previous studies. These include random forest (RF) [10], decision tree (DT) [11], support vector machine (SVM) [12], extra tree (ET) [13], light gradient-boosting machine (light GBM) [14], eXtreme gradient boosting (XGBoost) [15], AdaBoost [16], k-nearest neighbours (KNN) [17], naive Bayes (NB) [18], artificial neural network (ANN) [19] and multi-layer perceptron (MLP) [20]. The following subsections outline the review of (1) ML-based studies on crash severity prediction, (2) ML-based studies on DIS prediction and (3) ML and statistical-based studies on speeding-related crashes.

2.1 Machine learning-based studies on the crash severity prediction

In many studies, the predictive performance of the DT model was superior to other ML models in the crash injury severity prediction. For example, Cai et al. [21] reported that the DT model had better performance than the RF, logistic regression (LR), ANN and SVM models for crash severity prediction in the United States. This result was in agreement with the findings of Olutayo & Eludire [22] in Nigeria, Pakgohar et al. [23] in Iran and Taamneh et al. [24] in the United Arab Emirates.

Several past research studies conducted in the United States [25, 26], China [27, 28] and India [12, 29] reported the SVM model as the best-performing ML model compared to other statistical and ML models.

In China, Zhang et al. [30] compared the predictive performance of NB, KNN, XGBoost and binary logistic regression (BLR) models for the prediction of single-vehicle and multi-vehicle crash severity using five-year (2015–2019) data. They found that the XGBoost model outperformed the KNN, NB and BLR models in the injury severity prediction. The comparison results of the study were quite consistent with those reported by Mousa et al. [31] in the United States and Jamal et al. [32] in Saudi Arabia.

Many studies conducted in different countries have reported the superiority of the RF model over other ML models for crash severity prediction, such as Beshah et al. [33] in Ethiopia, Umer et al. [34] in the United States, Gang & Weng [35] in China, Wahab & Jiang [36] in Ghana, Chen & Chen [37] in Taiwan, Al-Mistarehi et al. [38] in Jordan and Bokaba et al. [10] in South Africa.

In China, Wei et al. [39] analysed and predicted the factors contributing to crash severity using six ML methods (light GBM, LR, SVM, KNN, XGBoost and DT). The result revealed that the light GBM model was considered the best ML model with accuracy = 92.6%, F1-score = 92.8% and the area under the curve (AUC) = 0.986. This result was consistent with the findings of [40] and [41]. Yang et al. [41] compared the

performance of the RF, light GBM and ANN models to identify the influential factors in driver-distracted crashes in the United States. They found that the light GBM model outperformed the other trained models.

The ANN, KNN, LR, ET, AdaBoost and MLP models were reported to be the top ML models for the prediction of crash injury severity in the studies of Alkheder et al. [19], Beshah & Hill [33], Fiorentini & Losa [42], Birfir et al. [13], Labib et al. [16] and Cicek et al. [5], respectively.

2.2 Machine learning-based studies on the driver injury severity prediction

In China, Zhou et al. [43] reported the XGBoost model as the superior model in the prediction of truck and car DIS when compared to the multinomial logit (MNL), NB, classification and regression tree (CART) and SVM models. A similar outcome was reported by Mokoatle et al. [15] in South Africa and Lee and Li [44] in South Korea. Mokoatle et al. [15] reported that the XGBoost model (above 80% accuracy) outperformed the multivariate logistic regression (MLR) model (around 60% accuracy).

The SVM-polynomial kernel model was the most accurate ML model for DIS prediction in a United States study [45]. The FIA results of the study revealed that driving under the influence of alcohol or drugs, seatbelt use, time of crash and crash location were significantly associated with DIS.

In the United States, Abdel-Aty & Abdelwahab [17] reported that the MLP model (accuracy = 73.5%) outperformed the fuzzy adaptive resonance theory neural network (70.6%) and ordered probit (OP) (61.7%) models in predicting DIS. Also, seat belt usage, driver gender, speed of vehicle, vehicle type, impact point and type of area were the influencing variables for the DIS.

Krishnaveni & Hemalatha [46] evaluated the predictive performance of NB, DT, AdaBoost and RF models for the DIS in India. The authors observed that the RF model showed a better performance among all trained ML models. By applying the RF model, Zhu et al. [47] examined the DIS patterns using the three-year Washington state data and found that truck vehicles, lack of restraint, female drivers, driver distraction and impairment, rollover type crash, time of crash and overtake operations were the main influential variables to the DIS in a single-vehicle run-off-road crash. In the United States, Mafi et al. [48] reported that the RF model showed superiority over the instance-based and DT models in the DIS prediction. A similar result was reported in a Spanish study [49], in which the authors found that the RF+CART model outperformed the SVM and binomial logit models in the prediction of the DIS.

Using police-reported six-year (1995–2000) crash data, Chong et al. [50] evaluated the performance of four ML models (DT, SVM, ANN and hybrid DT-ANN) for DIS prediction in the United States. They reported in their study that the hybrid DT-ANN model outperformed the SVM, ANN and DT models. The results of the CART indicated that vehicle type, drinking while driving, crash location, nature of crash, seatbelt use and number of vehicles involved were the main contributors to the DIS in truck crashes in Taiwan [51].

2.3 Machine learning and statistically-based studies on speeding-related crashes

In Taiwan, Tseng et al. [52] examined the contributory factors to truck DIS in speeding-related crashes using the BLR model. The study results implied that driver age and education, driving status and sleep quality factors were statistically significant in the speeding offences.

In Singapore, Zhou & Chin [53] investigated the various contributing factors to DIS in out-of-control single-vehicle crashes by applying the OP model. The results implied that age (≥ 65 years), drinking while driving, left and right turn manoeuvres and midnight driving were associated with higher rider and DIS.

Focusing on speeding-related crashes, Islam & Mannering [54] studied the gender and temporal instability effect in the DIS using a mixed logit model in Florida. The model results revealed that male and female drivers had significant differences in injury levels and that the effect of contributing factors to the DIS varied significantly over time.

In a cross-sectional study, Rashmi & Marisamynathan [55] examined the risk factors that lead to speeding-related crashes among long-haul truck Indian drivers using the BLR and ML models. The authors reported that the RF model demonstrated better performance in the prediction of speeding behaviour than the BLR, DT, AdaBoost and XGBoost models.

Zhang et al. [56] applied the geographically weighted neural network (GWNN) model to establish the relationships between contributing factors and injury severity of speeding-related crashes. They noted important spatial variability for several factors associated with injury severity in speeding-related crashes. These include driver condition, vehicle type, crash type, speed limit, weather, time and location of the crash, road configuration and volume of road traffic.

Previous studies have studied the influence of speeding (as an independent variable) on crash injury severity/driver injury severity using traditional statistical models [57–59].

From the above reviews, it is observed that many researchers have focused their studies on the analysis and prediction of crash severity or DIS using statistical models. Studies on speeding-related crashes were also carried out using these traditional methods. However, none of them have applied ML methods for the prediction of the DIS in speeding-related crashes. Moreover, most of these studies have been conducted in the United States and other developed nations, and limited literature from developing countries is available, particularly India. To address these gaps, the current study aims to predict DIS in speeding-related crashes by comparing the predictive performance of the ML models using a coding-free platform and analysing feature importance ratings of various contributing factors from the best ML model results using the SHAP approach.

3. DATA AND METHODS

3.1 Data and variables description

This study was conducted using a police-reported speeding-related crash dataset collected between 2011 and 2020 from Imphal, the capital of Manipur, India. According to the Imphal police report, a crash was classified as speeding-related if police officials confirmed that the crash occurred due to over-speeding (exceeding the assigned speed limit) or because the vehicle was out of control. After data pre-processing, a total of 793 single- and two-vehicle crash observations (individual drivers involved in a crash) were available for analysis and modelling. There were two levels of driver injury severity (DIS) for all crash observations in the dataset: non-fatal (not resulting in death within one month of the crash) and fatal (resulting in death within one month of the crash). Out of the 793 observations, 471 (68%) involved non-fatal injuries and 222 (32%) involved fatal injuries. The independent variables established from the dataset were driver age, gender, time of crash, vehicle type and nature of crash. The categorical variables of these five independent variables were “four age groups” (below 18, 18–24, 25–40 and above 40-year age group), “two genders” (male and female), “three vehicle types” (two-wheeler, light motor vehicle and heavy motor vehicle), “four times of crash” (12 AM–6 AM, 6 AM–12 PM, 12 PM–6 PM and 6 PM–12 AM), and “seven natures of crashes” (head-on collision, side-collision, rear collision, hit object, hit pedestrian, overturn and other). A statistical description of these categorical variables is provided in *Table 1*.

Table 1 – Descriptive statistics for categorical variables

Variables	Category	Non-fatal	Fatal	Total	Total (%)	Injury level (%)	
						Non-fatal	Fatal
Age (A)	Below 18	29	14	43	5.42	67.44	32.56
	18-24	131	40	171	21.56	76.61	23.39
	25-40	372	96	468	59.02	79.49	20.51
	Above 40	91	20	111	14.00	81.98	18.02
Gender (G)	Male	559	152	711	89.66	78.62	21.38
	Female	64	18	82	10.34	78.05	21.95
Vehicle type (VT)	Two-wheeler	243	95	338	42.62	71.89	28.11
	LMV	333	70	403	50.82	82.63	17.37
	HMV	47	5	52	6.56	90.38	9.62
Time of crash (T)	12 AM-6 AM	23	13	36	4.54	63.89	36.11
	6AM-12PM	125	22	147	18.54	85.03	14.97
	12 PM-6 PM	235	62	297	37.45	79.12	20.88
	6 PM-12 AM	240	73	313	39.47	76.68	23.32
Nature of crash (N)	Head-on collision	169	54	223	28.12	75.78	24.22
	Side collision	144	33	177	22.32	81.36	18.64
	Rear-end collision	128	26	154	19.42	83.12	16.88
	Hit object	64	24	88	11.10	72.73	27.27
	Hit pedestrian	64	18	82	10.34	78.05	21.95
	Overturn	38	14	52	6.56	73.08	26.92
	Other	16	1	17	2.14	94.12	5.88

3.2 Methods

The study flowchart for the prediction and investigation of DIS in speeding-related crashes is presented in Figure 1. Since the target variable (DIS) was binary (fatal and non-fatal), the classification approach was the most appropriate data mining technique for the present study. Eleven supervised ML algorithms used in this study were the gradient boosting tree (GBT), lasso-LARS, RF, XGBoost, DT, ET, stochastic gradient descent (SGD), KNN, light GBM, single-layer perceptron (SLP) and SVM (readers are requested to refer to the following article: Sorum & Pal [60] for a description of these algorithms). All ML algorithms were used with their default settings as available in the Dataiku platform, a coding-free platform for data science [61]. All variables in the dataset were transformed into integer values using the “Prepare” button in Dataiku. All the models were validated using stratified k-fold cross-validation (k-FCV), which is one of the most effective approaches, especially when dealing with unbalanced datasets. In this technique, the dataset was randomly divided into k folds of equal size, and the model was trained k times, with k-1 folds used for training and the remaining fold for validation. In this study, all models were implemented using 5- and 10-FCV with train ratios (TRs) of 0.5, 0.6, 0.7 and 0.8 in each FCV (*train ratio is defined as the percentage of the data that is allocated to the training set when the data are divided into a training set and a test set*). Five performance metrics (PMs)-area under the curve (AUC), F1 score, accuracy, recall and precision – were employed to compare the performance of the ML models for the DIS prediction. Accuracy, precision, recall and F1 score metrics can be determined using equations 1, 2, 3 and 4, respectively. The AUC represents the area under the receiver operating characteristic (ROC) curve with a maximum value of one, signifying a model with perfect classification.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where, TP = true positive, TN = true negative, FN = false negative and FP = false positive.

The FIA was carried out based on the top ML models of all four TR values (0.5, 0.6, 0.7 and 0.8) under 5- and 10-FCVs.

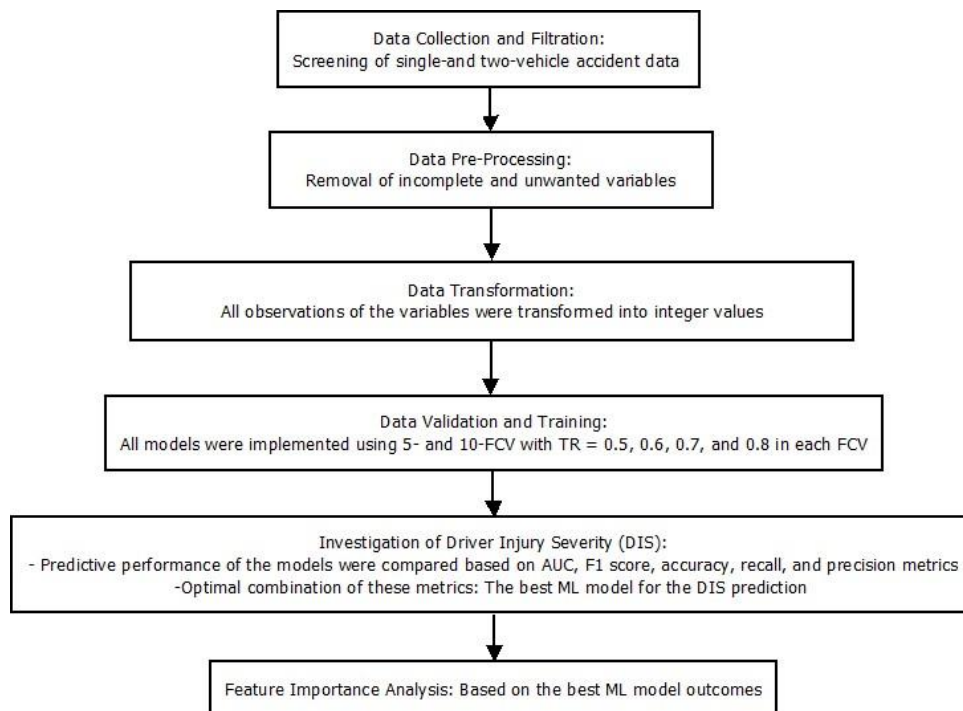


Figure 1 – Flowchart of the study methodology

4. RESULTS AND DISCUSSION

4.1 For 5-fold cross-validation

The PM values for TR = 0.5 are presented in Table 2. The SLP model is found to be the most accurate ML model among all trained models. More specifically, the accuracy of the SLP model is 60.3%, whereas the GBT model achieves the second-highest accuracy of 56.4%. The SGD model shows the lowest performance, with an accuracy of 22.1%. In terms of the precision metric, the results in Table 2 indicate that the GBT and SLP models provide the highest precision (28.6%) and outperform the remaining nine ML models. With the lowest precision value (22.1%), the SGD model is the poorest-performing model. The recall results (from Table 2) show that the XGBoost and SGD models (100%) achieve the best performance in DIS prediction, followed by the RF model, which has a recall value of 97.7%. The SLP model has the lowest recall value (53.5%) compared to the other trained models. The GBT model obtains the highest F1 score of 39.7%, followed by the RF model (39.3%). The SGD model ranks last in DIS prediction, with an F1 score of 36.1%. Based on the AUC metric, the GBT model (0.601) performs better overall than the other models, followed by the ET model with an AUC value of 0.587. The SGD model shows the worst predictive performance based on the AUC value.

Table 2 – Values of performance metrics at 5-FCV for TR = 0.5

SL.	ML algorithm	TR	Imphal at 5-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.5	33.3	24.6	97.7	39.3	0.586
2	GBT		56.4	28.6	65.1	39.7	0.601
3	Light GBM		38.2	23.9	82.6	37.1	0.556
4	XGBoost		29.7	23.9	100.0	38.6	0.579
5	DT		30.8	23.6	95.3	37.8	0.580
6	KNN		46.7	24.8	69.8	36.6	0.566
7	ET		34.6	24.6	95.3	39.1	0.587
8	SLP		60.3	28.6	53.5	37.2	0.549
9	SVM		46.9	25.3	72.1	37.5	0.569
10	Lasso-LARS		47.7	24.8	67.4	36.2	0.567
11	SGD		22.1	22.1	100.0	36.1	0.427

Note: Values of accuracy, precision, recall and F1 score are in %.

In terms of accuracy and precision metrics reported in Table 3, the SGD model performs well, obtaining 60.1% and 28.9%, followed by the GBT model (52.7% and 27.1%). The SLP model has very low accuracy and precision values (21.8%) compared to the other models. The recall results reveal that the SLP model exhibits the best performance for DIS prediction, with a recall value of 100%. The XGBoost model provides the second-highest recall value of 96.9%, while the KNN and SGD models generate the lowest recall value (56.9%). The light GBM model performs slightly better than the other models, as its F1 score is 39.5%, slightly higher than that of the GBT model (39.0%). The KNN model ranks last, with an F1 score of 31.8%. The GBT model achieves the highest AUC value of 0.619, followed by the SGD model (0.601). The KNN model, with an AUC value of 0.527, shows the worst performance for DIS prediction. In contrast, Beshah & Hill [33] reported that the KNN model was the most efficient model for the prediction of crash severity in Ethiopia.

Table 3 – Values of performance metrics at 5-FCV for TR = 0.6

SL.	ML algorithm	TR	Imphal at 5-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.6	38.9	24.9	89.2	38.9	0.576
2	GBT		52.7	27.1	69.2	39.0	0.619
3	Light GBM		36.2	24.9	95.4	39.5	0.594
4	XGBoost		32.2	24.0	96.9	38.4	0.575
5	DT		28.9	22.5	92.3	36.1	0.574
6	KNN		46.6	22.0	56.9	31.8	0.527
7	ET		36.6	24.6	92.3	38.8	0.579
8	SLP		21.8	21.8	100.0	35.8	0.536
9	SVM		47.3	25.5	73.8	37.9	0.583
10	Lasso-LARS		45.6	24.6	72.3	36.7	0.576
11	SGD		60.1	28.9	56.9	38.3	0.601

The study results reported in *Table 4* reveal that the DT model has the highest accuracy and precision values (69.1% and 31.1%), followed by the SGD model (56.4% and 26.2%). These findings are consistent with those reported by [50] and [62]. The lowest accuracy and precision values are produced by the XGBoost model (25% and 21.4%). Based on the recall metric, the RF and XGBoost models (100%) achieve the best results for DIS prediction. This aligns with the results reported by [36], which found that the RF model outperformed the DT and instance-based learning models in terms of recall and accuracy metrics. This was further confirmed by [35], which reported that the RF model had an advantage over the DT and AdaBoost models. The present study results indicate that the ET and DT models generate the second-highest and lowest recall values, at 97.8% and 42.2%, respectively. However, a previous study considered the ET model to be the best ML model for predicting bicyclist injury severity based on the recall metric [13]. The F1 score results indicate that the GBT model, with an F1 score of 37.8%, performs better than the other ML models, followed by the RF and lasso-LARS models (37.3%). The KNN model performs the worst, with the lowest F1 score of 33.3%. The AUC values reported in *Table 4* confirm the excellent classification performance of the lasso-LARS model (0.603).

Table 4 – Values of performance metrics at 5-FCV for TR = 0.7

SL.	ML algorithm	TR	Imphal at 5-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.7	31.4	23.0	100.0	37.3	0.581
2	GBT		37.3	23.7	93.3	37.8	0.598
3	Light GBM		44.5	23.8	77.8	36.5	0.575
4	XGBoost		25.0	21.4	100.0	35.3	0.512
5	DT		69.1	31.1	42.2	35.8	0.576
6	KNN		45.5	22.2	66.7	33.3	0.537
7	ET		31.8	22.8	97.8	37.0	0.563
8	SLP		49.5	24.2	68.9	35.8	0.568
9	SVM		44.1	23.3	75.6	35.6	0.561
10	Lasso-LARS		54.1	25.9	66.7	37.3	0.603
11	SGD		56.4	26.2	62.2	36.8	0.583

For the 5-FCV case with TR = 0.8, the highest accuracy value is associated with the lasso-LARS model (59.6%), indicating that it outperforms other models in DIS prediction (*Table 5*). The KNN model ranks second, with an accuracy of 42.9%. The poorest performance based on the accuracy metric is demonstrated by the SLP and SVM models (18%). In terms of precision, the lasso-LARS model again delivers the best performance, achieving the highest precision value of 23.5%, followed by the GBT model (20.6%). The lowest precision values are observed in the SVM and SLP models (18%). The recall results indicate that four ML models (RF, XGBoost, SLP and SVM) achieve a perfect recall of 100%. The ET model follows in second place with a recall of 96.6%, while the lasso-LARS model ranks last with 55.2%. The highest F1 score (33.5%) is shared by the GBT and light GBM models, followed closely by the RF model (33.3%). The SGD model demonstrates the best performance based on the AUC metric, achieving a value of 0.597. The second-best AUC score is shared by the lasso-LARS and light GBM models (0.546), which slightly outperform the DT model (0.528).

Table 5 – Values of performance metrics at 5-FCV for TR = 0.8

SL.	ML algorithm	TR	Imphal at 5-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.8	28.0	20.0	100.0	33.3	0.511
2	GBT		36.0	20.6	89.7	33.5	0.517
3	Light GBM		33.5	20.5	93.1	33.5	0.546
4	XGBoost		25.5	19.5	100.0	32.6	0.482
5	DT		26.7	19.3	96.6	32.2	0.528
6	KNN		42.9	18.2	62.1	28.1	0.494
7	ET		28.6	19.7	96.6	32.7	0.512
8	SLP		18.0	18.0	100.0	30.5	0.492
9	SVM		18.0	18.0	100.0	30.5	0.457
10	Lasso-LARS		59.6	23.5	55.2	33.0	0.546
11	SGD		31.7	19.5	89.7	32.1	0.597

Overall, from the above discussion, it is observed that the GBT model provides the best combination of PMs and can be considered the top ML model for DIS prediction when TR = 0.5 under 5-FCV. Similarly, for TR = 0.6, 0.7 and 0.8, the SGD, DT and lasso-LARS models, respectively, can be selected as the top ML models for DIS prediction in Imphal speeding-related crashes. These findings are summarised in *Table 6*.

Table 6 – Summary of the top model performance for 5-FCV

k-fold	TR	Top model	Performance metrics				
			Accuracy	Precision	Recall	F1 score	AUC
5	0.5	GBT	56.4	28.6	65.1	39.7	0.601
	0.6	SGD	60.1	28.9	56.9	38.3	0.601
	0.7	DT	69.1	31.1	42.2	35.8	0.576
	0.8	Lasso-LARS	59.6	23.5	55.2	33.0	0.546

4.2 For 10-fold cross-validation

The PM values for TR = 0.5 under 10-FCV are tabulated in *Table 7*. The SLP model achieves the highest accuracy (60.3%) among all trained ML models, making it the most accurate predictive model. This is followed by the GBT model, with an accuracy of 56.4%. The XGBoost model (28.7%) ranks last in the comparison. The GBT and SLP models (28.6%) exhibit the best performance in terms of precision. The SGD model ranks second, with a precision value of 26.2%, indicating strong predictive performance. The lowest precision values are observed in two models (XGBoost and DT). However, the XGBoost model achieves a 100% recall value, followed by the RF model (97.7%), which aligns with the findings of [44], [15] and [18]. The SLP model demonstrates the worst predictive performance in terms of recall, with a value of 53.5%. The F1 score results reported in *Table 7* indicate that the light GBM model provides the highest F1 score (40.2%) among all other models. Wen et al. [11] reported a similar finding in their study, stating that the light GBM outperformed the XGBoost. The GBT model ranks second, with a slightly higher F1 score (39.7%) than the RF (39.3%) and ET (39.1%) models. The lasso-LARS model has the lowest F1 score (36.2%). Furthermore, all trained ML models showed reasonable prediction ability in terms of AUC, with the light GBM covering the largest area (0.620). The second-highest and lowest areas covered under the ROC curve are associated with the SGD (0.602) and SLP (0.549) models, respectively.

Table 7 – Values of performance metrics at 10-FCV for TR 0.5

SL.	ML algorithm	TR	Imphal at 10-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.5	33.3	24.6	97.7	39.3	0.586
2	GBT		56.4	28.6	65.1	39.7	0.601
3	Light GBM		37.4	25.5	95.3	40.2	0.620
4	XGBoost		28.7	23.6	100.0	38.2	0.566
5	DT		30.8	23.6	95.3	37.8	0.580
6	KNN		46.7	24.8	69.8	36.6	0.566
7	ET		34.6	24.6	95.3	39.1	0.587
8	SLP		60.3	28.6	53.5	37.2	0.549
9	SVM		47.7	25.4	70.9	37.4	0.569
10	Lasso-LARS		47.7	24.8	67.4	36.2	0.567
11	SGD		49.0	26.2	72.1	38.4	0.602

The comparison results from *Table 8* show that the SVM model has the highest accuracy and precision values (62.8% and 29.1%), followed by the SGD model (60.1% and 28.9%). A similar result was reported by [42]. The SLP model demonstrates the poorest predictive performance with the lowest accuracy (21.8%) and precision (21.8%). In terms of the recall metric, the XGBoost and SLP models exhibit the best performance with a 100% value. These are followed by the DT and ET models (92.3%). The SVM model has the lowest recall value (49.2%). The highest F1 score value (39.0%) is shared by the GBT and light GBM models, indicating them as the most efficient predictive-performing models. The RF and KNN models occupy the

second and last positions, respectively, with F1 score values of 38.9% and 31.8%. In comparison, the GBT model demonstrates the highest predictive capability, with an AUC value of 0.619, followed by the light GBM model (AUC = 0.603). The KNN model covers the minimum area under the ROC curve (AUC = 0.527).

Table 8 – Values of performance metrics at 10-FCV for TR 0.6

SL.	ML algorithm	TR	Imphal at 10-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.6	38.9	24.9	89.2	38.9	0.576
2	GBT		52.7	27.1	69.2	39.0	0.619
3	Light GBM		49.7	26.5	73.8	39.0	0.603
4	XGBoost		27.9	23.2	100.0	37.7	0.565
5	DT		28.9	22.5	92.3	36.1	0.574
6	KNN		46.6	22.0	56.9	31.8	0.527
7	ET		36.6	24.6	92.3	38.8	0.579
8	SLP		21.8	21.8	100.0	35.8	0.536
9	SVM		62.8	29.1	49.2	36.6	0.583
10	Lasso-LARS		45.6	24.6	72.3	36.7	0.576
11	SGD		60.1	28.9	56.9	38.3	0.601

In terms of accuracy and precision metrics reported in Table 9, the DT model achieves the highest accuracy and precision values (69.1 % and 31.1%), followed by the SGD model (56.4% and 26.2%). The XGBoost model, with accuracy = 25.0% and precision = 21.4%, ranks last in DIS prediction. However, the same model (XGBoost), along with the RF model, attains the highest recall value (100%). The second-highest and lowest recall values are recorded by the ET and DT models (97.8% and 42.2%), respectively. The light GBM model (39.6% and 0.615) performs better than other ML models in terms of the F1 score and AUC values. This is followed by the GBT (w.r.t the F1 score = 37.8%) and the lasso-LARS (w.r.t the AUC = 0.603). The KNN model exhibits the lowest F1 score and AUC values (33.3% and 0.537) for TR = 0.7 under 10-FCV.

Table 9 – Values of performance metrics at 10-FCV for TR 0.7

SL.	ML algorithm	TR	Imphal at 10-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.7	31.4	23.0	100.0	37.3	0.581
2	GBT		37.3	23.7	93.3	37.8	0.598
3	Light GBM		41.8	25.1	93.3	39.6	0.615
4	XGBoost		25.0	21.4	100.0	35.3	0.512
5	DT		69.1	31.1	42.2	35.8	0.576
6	KNN		45.5	22.2	66.7	33.3	0.537
7	ET		31.8	22.8	97.8	37.0	0.563
8	SLP		49.5	24.2	68.9	35.8	0.568
9	SVM		46.4	24.1	75.6	36.6	0.561
10	Lasso-LARS		54.1	25.9	66.7	37.3	0.603
11	SGD		56.4	26.2	62.2	36.8	0.583

The values of PMs for TR = 0.8 under 10-FCV are presented in Table 10. With a 59.6% accuracy value, the lasso-LARS model provides the best performance for the DIS prediction, followed by the KNN model (42.9%). The SLP and SVM models perform the worst for the DIS prediction, with an accuracy value of 18.0%. The lasso-LARS model achieves the highest precision value (23.5%), followed by the light GBM model (20.9%). The SLP and SVM models stand in the last position with the lowest precision value (18%). However, the same models (SLP and SVM), along with the RF and XGBoost, achieve the highest recall value (100%). The second-highest recall value is recorded by the DT and ET (96.6%), while the lasso-LARS has the lowest recall value (55.2%). Based on the F1 score metric, the light GBM model (34.2%) demonstrates the best performance for

DIS prediction, followed by the GBT model (33.5%). The KNN model ranks last, with an F1 score of 28.1%. The best DIS prediction, based on the AUC value, is exhibited by the SGD model. The second-highest area covered under the curve is associated with the light GBM model (0.556).

Table 10 – Values of performance metrics at 10-FCV for TR 0.8

SL.	ML algorithm	TR	Imphal at 10-FCV				
			Accuracy	Precision	Recall	F1 score	AUC
1	RF	0.8	28.0	20.0	100.0	33.3	0.511
2	GBT		36.0	20.6	89.7	33.5	0.517
3	Light GBM		35.4	20.9	93.1	34.2	0.556
4	XGBoost		23.0	19.0	100.0	31.9	0.457
5	DT		26.7	19.3	96.6	32.2	0.528
6	KNN		42.9	18.2	62.1	28.1	0.494
7	ET		28.6	19.7	96.6	32.7	0.512
8	SLP		18.0	18.0	100.0	30.5	0.492
9	SVM		18.0	18.0	100.0	30.5	0.457
10	Lasso-LARS		59.6	23.5	55.2	33.0	0.546
11	SGD		31.7	19.5	89.7	32.1	0.597

Overall, it is observed that the light GBM model provides the best combination of PMs and can be considered the top ML model for DIS prediction when TR = 0.5 and 0.7 under 10-FCV. Similarly, for TR = 0.6 and 0.8, the GBT and lasso-LARS models, respectively, can be selected as the top ML models for DIS prediction in Imphal speeding-related crashes. These results are summarised in Table 11.

Table 11 – Summary of the top model performance for 10-FCV

k-fold	TR	Top model	Performance metrics				
			Accuracy	Precision	Recall	F1 score	AUC
10	0.5	Light GBM	37.4	25.5	95.3	40.2	0.620
	0.6	GBT	52.7	27.1	69.2	39.0	0.619
	0.7	Light GBM	41.8	25.1	93.3	39.6	0.615
	0.8	Lasso-LARS	59.6	23.5	55.2	33.0	0.546

4.3 Feature importance analysis

Feature importance analysis (FIA) based on Shapley values is referred to as Shapley Additive exPlanations (SHAP). SHAP values illustrate how each attribute contributes to the prediction for each occurrence. A high SHAP value represents the feature as significantly more important. It is important to note that the FIA using SHAP was performed only on the best ML models for TR values of 0.5, 0.6, 0.7 and 0.8 under 5- and 10-FCVs (as summarised in Tables 6 and 11).

The FIA results for 5-FCV from Figure 2 (a-d) revealed that the two-wheeler (17.11%), head-on collision (23.75%), 6 AM–12 PM (21.06%) and 12 PM–6 PM (14.57%) variables had the maximum impact on the DIS prediction for TR values of 0.5, 0.6, 0.7 and 0.8, respectively. These were followed by the 6 AM–12 PM (10.82%), 12 PM–6 PM (13.53%), hit-pedestrian (19.86%) and 6 PM–12 AM (13.64%) variables for TR = 0.5, 0.6, 0.7 and 0.8, respectively. Female drivers (1.07%), LMV (0.15%), above 40 age group (0.30%) and other (0.03%) variables were found to have the minimum impact on the DIS prediction for TR values of 0.5, 0.6, 0.7 and 0.8, respectively. It was further noticed that the gender variable had no impact on the DIS prediction in most of the cases. A similar outcome was reported by [33], [35] and [42]. The authors reported that the time of crash and nature of crash (collision type) were the most critical determinants of crash severity. However, some previous studies reported different outcomes. For instance, Chen et al. [45] reported that the HMV variable was the most influential factor for the DIS when a hybrid DT and NB model was deployed. Another study reported that the LMV variable was the top contributor to the DIS when the XGBoost model

was deployed [15]. In a recent study, it was reported that the driver age variable was one of the most influential factors for predicting speeding behaviours among long-haul truck drivers in India [55]. The results of this study implied that the feature and its percentage impact on the DIS prediction in FIA depended upon the TR value and indirectly depended on the ML models deployed. For example, the position and percentage impact of the 12 PM-6 PM variable were 5th (6.02%), 2nd (13.53%), 4th (7.7%) and 1st (14.57%) when TR = 0.5 (GBT model), = 0.6 (SGD model), = 0.7 (DT model) and = 0.8 (lasso-LARS model).

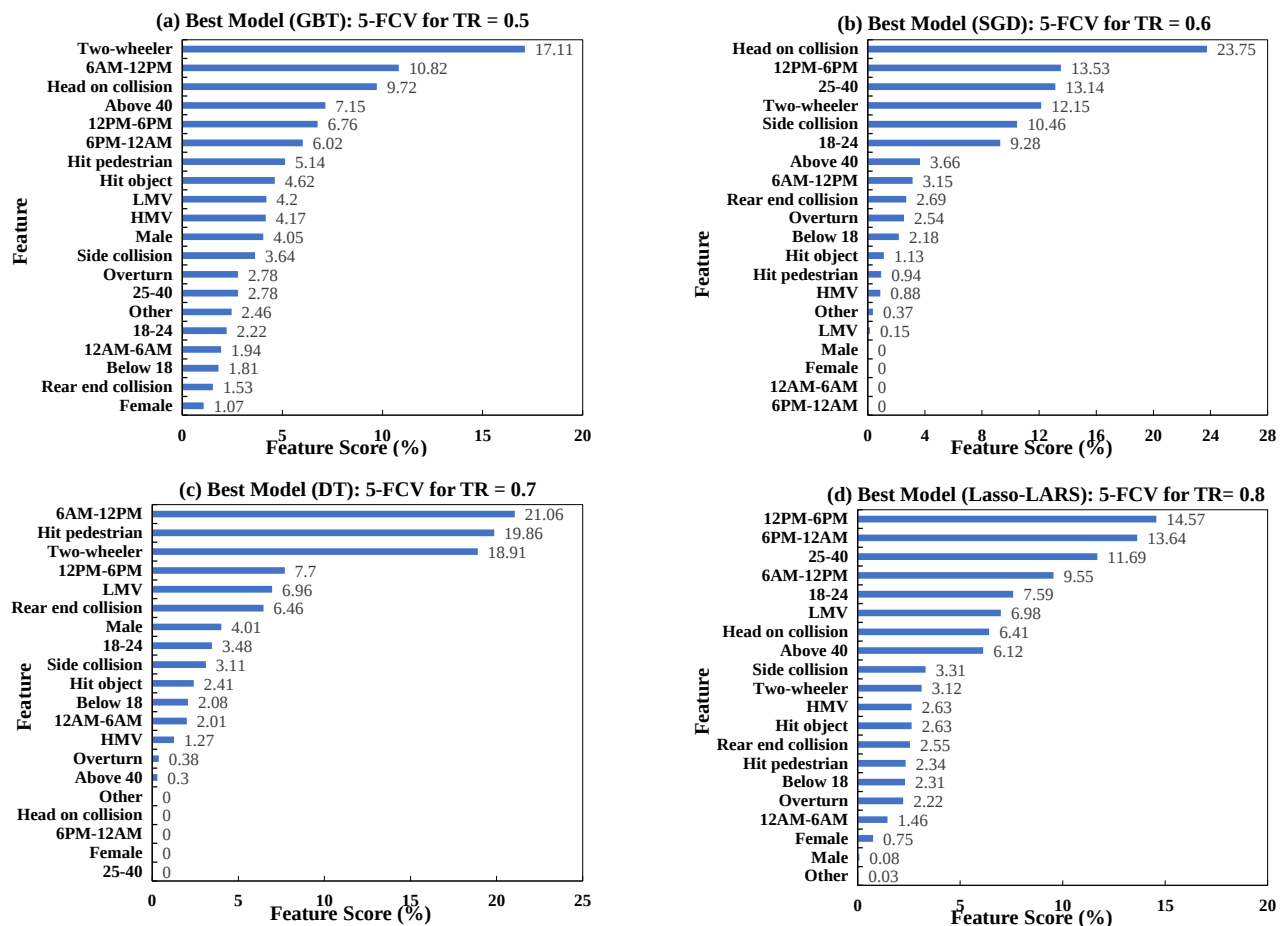


Figure 2 – Feature importance rating for 5-FCV when (a) TR = 0.5, (b) TR = 0.6, (c) TR = 0.7 and (d) TR = 0.8

The FIA results for 10-FCV from Figure 3 (a-d) indicated that the two-wheeler (18.20%), head-on collision (16.44%), two-wheeler (30.31%) and 12 PM–6 PM (13.1%) variables contributed the highest impact on the DIS prediction for TR values of 0.5, 0.6, 0.7 and 0.8, respectively. The second most impactful variables were the 6 AM–12 PM (15.43%), 12 PM–6 PM (12.39%), head-on collision (21.83%) and 25–40 age group (12.19%) in the case of TR = 0.5, 0.6, 0.7 and 0.8, respectively. The least impactful variables in the case of TR = 0.5, 0.6, 0.7 and 0.8 were the overturn (0.22%), 6 PM–12 AM (1.3%), rear-end collision (0.68%) and male drivers (0.19%), respectively. However, Mokoatle et al. [15] stated that overturned-type crashes were one of the most contributing variables for increased DIS. Furthermore, the result of this study indicated that the “other” variable had no impact on the DIS prediction (except in TR = 0.6). Similar to the 5-FCV case, the results of the 10-FCV case also showed that the feature and its percentage impact on the DIS prediction in FIA depended upon the TR value and ML models deployed.

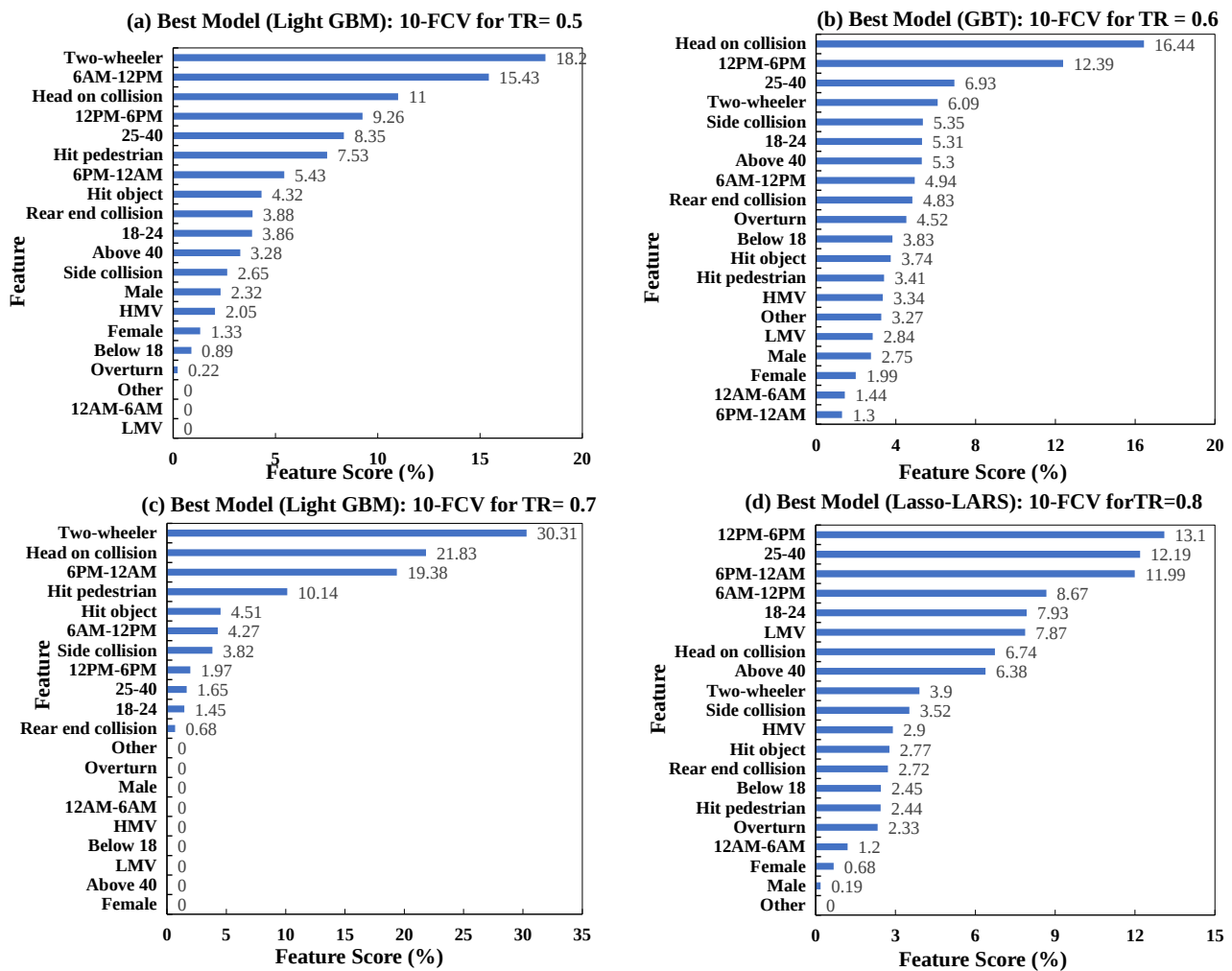


Figure 3 – Feature importance rating for 10-FCV when (a) TR = 0.5, (b) TR = 0.6, (c) TR = 0.7 and (d) TR = 0.8

Overall, it can be concluded that the vehicle type (two-wheeler), nature of crash (head-on collision), and time of crash (12 PM–6 PM and 6 AM–12 PM) variables have significant impacts on the DIS prediction in Imphal speeding-related crashes. This overall outcome is in line with the results reported by [59]. Further, it is observed that the degree of variable impact on the DIS prediction varies with the change in FCV value for the same TR value. For instance, the percentage impact of the two-wheeler and 6 AM–12 PM variables for TR =0.5 differ between 5- and 10-FCV cases.

5. SUMMARY AND CONCLUSIONS

The present study aimed to predict and analyse DIS in speeding-related crashes by applying eleven supervised ML algorithms on the Dataiku platform using 793 police-reported single- and two-vehicle crash data collected from Imphal City between 2011 and 2020. For DIS prediction, all ML models were implemented using 5- and 10-FCVs and trained at TR values of 0.5, 0.6, 0.7 and 0.8 in each FCV. While comparing predictive performance, the top ML model for DIS prediction was selected based on the best combination of recall, precision, F1 score, accuracy and AUC metrics. From the top ML models, the importance of variables contributing to DIS in speeding-related crashes was determined through FIA. The important conclusions drawn from the study are summarised below.

– For the 5-FCV, the GBT, SGD, DT and lasso-LARS models were identified as the top ML models for the DIS prediction in speeding-related crashes for TR = 0.5, 0.6, 0.7 and 0.8, respectively. For the 10-FCV, the light GBM model was the top ML model for the DIS prediction in speeding-related crashes for TR = 0.5 and 0.7 cases. In the case of TR = 0.6 and 0.8, the GBT and lasso-LARS models, respectively, were the top ML-performing models for the DIS prediction.

- The feature and its percentage impact on the DIS prediction in FIA depended upon the TR value and indirectly depended on the ML models deployed.
- The degree of variable impact on the DIS prediction varies with the changes in FCV value for the same TR value
- Vehicle type (two-wheeler), nature of crash (head-on collision) and time of crash (12 PM–6 PM and 6 AM–12 PM) variables were the most impactful variables on the DIS prediction in Imphal speeding-related crashes.

These ML models can be employed in hilly areas for the accurate prediction of DIS. The results of the current study may be helpful for transportation policy-makers to have a better knowledge of various contributing elements to the RTAs while implementing road safety measures to reduce speeding-related crash severities for drivers. This study has some limitations. Firstly, since the police-reported data were used, there is a possibility of under-reporting some crash data and missing other injury severity levels. Secondly, the small size of the crash data was used, which may lead to biased results. Additionally, the unobserved heterogeneity in crash severity modelling was not considered in the present study. To address these issues, the following recommendations are proposed:

- To overcome the limitation of under-reporting in police-reported data and also the missing severity level of injuries, future studies should consider supplementing additional data from other sources, such as hospitals or insurance companies. This will help provide more comprehensive and accurate information regarding crash severity and injuries.
- For small sample size issue, future studies should rely more on larger datasets by merging data from multiple regions to represent a wider spectrum of crash types over a broader range of severity levels, or use advanced statistical methods, such as bootstrapping or Bayesian approaches, to better utilise data and reduce possible biases.
- Researchers should consider implementing more advanced modelling approaches to control for unobserved heterogeneity. One possibility is hybrid approaches, combining ML with statistical models (e.g. mixed logit models or generalised linear models) to reduce the impact of unobserved heterogeneity.

ACKNOWLEDGEMENTS

The authors would like to extend their gratitude towards the Superintendents of Police (Traffic Cell), Department of Police, Imphal, India, for providing crash data of the city.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The data analysed in this article will be made available upon request to the corresponding author.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

REFERENCES

- [1] World Health Organization (WHO). Global status report on road safety 2018: Summary 2018. *World Health Organization*, 2018.
- [2] MoRTH. Road Accidents in India 2021. *Ministry of road transport and highway*. 2021;1–237. DOI: 10.1016/s0386-1112(14)60239-9.

- [3] Ye F, et al. Investigating the severity of expressway crash based on the random parameter logit model accounting for unobserved heterogeneity. *Advances in Mechanical Engineering*. 2021;13:1–13. DOI: 10.1177/16878140211067278.
- [4] Guo Y, Wu Y, Lu J, Zhou J. Modeling the unobserved heterogeneity in E-bike collision severity using full bayesian random parameters multinomial logit regression. *Sustainability*. 2019;11. DOI: 10.3390/su11072071.
- [5] Mannering FL, Shankar V, Bhat CR. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods Accident Research*. 2016;11:1–16. DOI: 10.1016/j.amar.2016.04.001.
- [6] Tang J, et al. Crash injury severity analysis using a two-layer Stacking framework. *Accident Analysis Prevention*. 2019;122:226–38. DOI: 10.1016/j.aap.2018.10.016.
- [7] Se C, et al. Modeling motorcycle crash-injury severity utilizing explainable data-driven approaches. *Transportation Letters*. 2024;00:1–26. DOI: 10.1080/19427867.2024.2408920.
- [8] Cicek E, Akin M, Uysal F, Topcu Aytas RM. Comparison of traffic accident injury severity prediction models with explainable machine learning. *Transportation Letters*. 2023;15:1043–54. DOI: 10.1080/19427867.2023.2214758.
- [9] Lundberg SM, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*. 2018;2:749–60. DOI: 10.1038/s41551-018-0304-0.
- [10] Bokaba T, Doorsamy W, Paul BS. Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Science*. 2022;12. DOI: 10.3390/app12020828.
- [11] Cai X, Rahmati Y, Jain S, Fishelson J. Machine learning methods to analyze and predict crash injury severity based on contributing factors for southeast Michigan. *Transportation Research Record*. 2023;2677:83–94. DOI: 10.1177/03611981221113569.
- [12] Jha AN, Chatterjee N, Tiwari G. A performance analysis of prediction techniques for impacting vehicles in hit-and-run road accidents. *Accident Analysis Prevention*. 2021;157:106164. DOI: 10.1016/j.aap.2021.106164.
- [13] Birfir S, Elalouf A, Rosenbloom T. Building machine-learning models for reducing the severity of bicyclist road traffic injuries. *Transportation Engineering*. 2023;12:100179. DOI: 10.1016/j.treng.2023.100179.
- [14] Wen X, Xie Y, Wu L, Jiang L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accident Analysis Prevention*. 2021;159. DOI: 10.1016/j.aap.2021.106261.
- [15] Mokoatle M, Marivate V, Esiefarienrhe M. Predicting road traffic accident severity using accident report data in South Africa. *ACM 20th Annual International Conference on Digital Government Research, June 18–20, 2019, Dubai, United Arab Emirates*. 2019:11–7. DOI: 10.1145/3325112.3325211.
- [16] Labib MF, et al. Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. *7th International Conference on Smart Computing & Communications (ICSCC), 28-30 June, 2019, Sarawak, Malaysia*. 2019:1–5. DOI: 10.1109/ICSCC.2019.8843640.
- [17] Zhang S, et al. Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. *PLoS One*. 2022;17:1–19. DOI: 10.1371/journal.pone.0262941.
- [18] Zhou B, et al. Comparing factors affecting injury severity of passenger car and truck drivers. *IEEE Access*. 2020;8:153849–61. DOI: 10.1109/ACCESS.2020.3018183.
- [19] Alkheder S, Taamneh M, Taamneh S. Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*. 2017;36:100–8. DOI: 10.1002/for.2425.
- [20] Abdel-Aty MA, Abdelwahab HT. Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering*. 2004;130:204–10. DOI: 10.1061/(ASCE)0733-947X(2004)130:2(204).
- [21] Cai X, Rahmati Y, Jain S, Fishelson J. Machine learning methods to analyze and predict crash injury severity based on contributing factors for southeast Michigan. *Transportation Research Record*. 2023;2677:83–94. DOI: 10.1177/03611981221113569.
- [22] Olutayo AA, Eludire AA. Traffic accident analysis using decision trees and neural networks. *International Journal of Information Technology Computer Science*. 2014;6:22–8. DOI: 10.5815/ijitcs.2014.02.03.
- [23] Pakgohar A, Tabrizi RS, Khalili M, Esmaeili A. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. *Procedia Computer Science*. 2011;3:764–9. DOI: 10.1016/j.procs.2010.12.126.
- [24] Taamneh M, Alkheder S, Taamneh S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety & Security*. 2017;9:146–66. DOI: 10.1080/19439962.2016.1152338.
- [25] Delen D, Tomak L, Topuz K, Eryarsoy E. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*. 2017;4:118–31. DOI: 10.1016/j.jth.2017.01.009.

- [26] Mokhtarimousavi S, Anderson JC, Azizinamini A, Hadi M. Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transportation Research Record*. 2019;2673:680–92. DOI: 10.1177/0361198119845899.
- [27] Liao Y, et al. Study on crash injury severity prediction of autonomous vehicles for different emergency decisions based on support vector machine model. *Electronics*. 2018;7. DOI: 10.3390/electronics7120381.
- [28] Li Z, Liu P, Wang W, Xu C. Using support vector machine models for crash injury severity analysis. *Accident Analysis Prevention*. 2012;45:478–86. DOI: 10.1016/j.aap.2011.08.016.
- [29] Priyanka A, Sathiyakumari. A comparative study of classification algorithm using accident data. *International Journal of Computer Science & Engineering Technology*. 2014;3:1018–23.
- [30] Zhang S, et al. Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. *PLoS One*. 2022;17:1–19. DOI: 10.1371/journal.pone.0262941.
- [31] Mousa SR, Bakhit PR, Ishak S. An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: A naturalistic driving study. *Canadian Journal of Civil Engineering*. 2019;46:712–21. DOI: 10.1139/cjce-2018-0117.
- [32] Jamal A, et al. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International Journal of Injury Control and Safety Promotion*. 2021;28:408–27. DOI: 10.1080/17457300.2021.1928233.
- [33] Beshah T, Hill S. Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. *AAAI Conference and Symposium Proceedings*. 2010;SS-10-01:14–9.
- [34] Umer M, Sadiq S, Sadiq A, Ullah S. Comparison analysis of tree based and ensembled regression algorithms for traffic accident and severity prediction. *ArXiv, Abs/201014921*. 2020:33–7.
- [35] Gan X, Weng J. Predicting crash injury severity for the highways involving traffic hazards and those involving no traffic hazards. *CICTP 2020 Advanced Transportation Technologies and Development-Enhancing Connections - Proceedings 20th COTA International Conference of Transportation Professionals*. 2020:4195–206. DOI: 10.1061/9780784482933.360.
- [36] Wahab L, Jiang H. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS One*. 2019;14:1–17. DOI: 10.1371/journal.pone.0214966.
- [37] Chen MM, Chen MC. Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. *Information*. 2020;11. DOI: 10.3390/INFO11050270.
- [38] Al-Mistarehi B, Alomari AH, Imam R, Mashaqba M. Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS. *Frontiers in Built Environment*. 2022;8:1–14. DOI: 10.3389/fbuil.2022.860805.
- [39] Wei T, Zhu T, Lin M, Liu H. Predicting and factor analysis of rider injury severity in two-wheeled motorcycle and vehicle crash accidents based on an interpretable machine learning framework. *Traffic Injury Prevention*. 2024;25:194–201. DOI: 10.1080/15389588.2023.2284111.
- [40] Wen X, Xie Y, Wu L, Jiang L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accident Analysis Prevention*. 2021;159. DOI: 10.1016/j.aap.2021.106261.
- [41] Yang C, et al. A machine learning approach to understanding the road and traffic environments of crashes involving driver distraction and inattention (DDI) on rural multilane highways. *Journal of Safety Research*. 2025;92:14–26. DOI: 10.1016/j.jsr.2024.11.011.
- [42] Fiorentini N, Losa M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*. 2020;5. DOI: 10.3390/infrastructures5070061.
- [43] Zhou D, Gayah VV., Wood JS. Integration of machine learning and statistical models for crash frequency modeling. *Transportation Letters*. 2022;00:1–12. DOI: 10.1080/19427867.2022.2158257.
- [44] Lee C, Li X. Predicting driver injury severity in single-vehicle and two-vehicle crashes with boosted regression trees. *Transportation Research Record*. 2015;2514:138–48. DOI: 10.3141/2514-15.
- [45] Chen C, et al. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis Prevention*. 2016;90:128–39. DOI: 10.1016/j.aap.2016.02.011.
- [46] Krishnaveni S, Hemalatha M. A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications*. 2011;23:40–8. DOI: 10.5120/2896-3788.
- [47] Zhu M, Li Y, Wang Y. Design and experiment verification of a novel analysis framework for recognition of driver injury patterns: From a multi-class classification perspective. *Accident Analysis Prevention*. 2018;120:152–64. DOI: 10.1016/j.aap.2018.08.011.

- [48] Mafi S, AbdelRazig Y, Doczy R. Machine learning methods to analyze injury severity of drivers from different age and gender groups. *Transportation Research Record*. 2018;2672:171–83. DOI: 10.1177/0361198118794292.
- [49] Pillajo-Quijia G, Arenas-Ramírez B, González-Fernández C, Aparicio-Izquierdo F. Influential factors on injury severity for drivers of light trucks and vans with machine learning methods. *Sustainability*. 2020;12. DOI: 10.3390/su12041324.
- [50] Chong M, Abraham A, Paprzycki M. Traffic accident analysis using machine learning paradigms. *Informatica*. 2005;29:89–98.
- [51] Chang LY, Wang HW. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis Prevention*. 2006;38:1019–27. DOI: 10.1016/j.aap.2006.04.009.
- [52] Tseng CM, et al. A comprehensive analysis of factors leading to speeding offenses among large-truck drivers. *Transportation Research Part F: Traffic Psychology Behaviour*. 2016;38:171–81. DOI: 10.1016/j.trf.2016.02.007.
- [53] Zhou M, Chin HC. Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. *Accident Analysis Prevention*. 2019;124:104–12. DOI: 10.1016/j.aap.2019.01.009.
- [54] Islam M, Mannering F. The role of gender and temporal instability in driver-injury severities in crashes caused by speeds too fast for conditions. *Accident Analysis Prevention*. 2021;153. DOI: 10.1016/j.aap.2021.106039.
- [55] Rashmi BS, Marisamynathan S. Investigating the contributory factors influencing speeding behavior among long-haul truck drivers traveling across India: Insights from binary logit and machine learning techniques. *International Journal of Transportation Science & Technology*. 2024;1–18. DOI: 10.1016/j.ijtst.2024.01.008.
- [56] Zhang Z, Xu N, Liu J, Jones S. Exploring spatial heterogeneity in factors associated with injury severity in speeding-related crashes: An integrated machine learning and spatial modeling approach. *Accident Analysis Prevention*. 2024;206:107697. DOI: 10.1016/j.aap.2024.107697.
- [57] Rifaat SM, Tay R, De Barros AG. Logistic model of injury risks in single vehicle crashes in urban neighborhoods. *Journal of Advanced Transportation*. 2011;45:186–95. DOI: 10.1002/atr.164.
- [58] Haleem K, Alluri P, Gan A. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accident Analysis Prevention*. 2015;81:14–23. DOI: 10.1016/j.aap.2015.04.025.
- [59] Kim J, Ulfarsson GF, Kim S, Shankar VN. Driver-injury severity in single-vehicle crashes in California : A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis Prevention*. 2013;50:1073–81. DOI: 10.1016/j.aap.2012.08.011.
- [60] Sorum NG, Pal D. Identification of the best machine learning model for the prediction of driver injury severity. *International Journal of Injury Control and Safety Promotion*. 2024;0:1–16. DOI: 10.1080/17457300.2024.2335478.
- [61] Dataiku 2014. Pronouncing Dataiku. <https://blog.dataiku.com/2014/08/07/pronouncing-dataiku>. (Accessed 12 April 2023). Dataiku, 2014 2014.
- [62] Chang L, Chien J. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Safety Science*. 2013;51:17–22. DOI: 10.1016/j.ssci.2012.06.017.