Prediction of Unknown Striking Vehicles in Motorized Two-Wheeler Hit-and-Run Crashes in Delhi

Anurag THOMBRE[a], Jatin BHANDARI[b], Amit AGARWAL[c], Indrajit GHOSH[d]

[a,b,c,d] *Indian Institute of Technology (IIT) Roorkee, Roorkee, 247667, India*
[a]*E-mail:* ***athombre@ce.iitr.ac.in***
[b]*E-mail:* ***jatin_b@mfs.iitr.ac.in***
[c]*E-mail:* ***amitfce@iitr.ac.in***
[d]*E-mail:* ***indrafce@iitr.ac.in***

Abstract: As the nations embark into the second decade of action for road safety, it is opportune that we critically review past mistakes and emphasize thrust areas to meet road safety targets. Road safety of vulnerable road users (VRUs) and hit-and-run road crashes are two areas with alarming trends in the past decade and necessitate concerted efforts. India, as the world leader in road traffic fatalities, is observing threatening numbers of VRUs and hit-and-run road crashes. The present study focuses on providing a solution to these correlated road safety issues by predicting the unknown striking vehicle type in case of hit-and-run road crashes involving motorized two-wheelers as the victim. Delhi, the capital of India, is the study area for the experiment. Predictive techniques such as supervised learning classification models are employed. Ensemble learning technique, such as Random Forest, has been found to perform best and have the maximum capability to predict the unknown striking vehicle type in hit-and-run road crashes involving motorized two-wheelers. The study findings are helpful for traffic enforcement agencies and policymakers to strategize action and execute prevention plans to improve the overall road safety situation.

Keywords: Motorized Two-Wheeler, Hit-and-Run Crash, Striking Vehicle, Road Safety Delhi

1. INTRODUCTION

Transportation has been an enabler for human beings in many ways in improving the quality of life. However, the need for travel to accomplish daily activities has also negatively impacted human health. Road crashes are one of the prime contributors (IHME, 2017). The heavy burden of road crashes can be understood by the fact that approximately 1.3 million people lose their lives every year, and between 20 and 50 million more people are suffering from road crash-related injuries of different severities, with many disabled for the rest of their lives (WHO, 2018).

Understanding the causal factors behind road crashes is vital in mitigating this mammoth global problem. However, one of the main impediments in determining the cause of road crashes is not having information about the offender's vehicle or the striking vehicle in a particular road crash. Typically, the offender's vehicle is considered as one that flees from the road site post committing the crash. These vehicles are commonly named 'unknown' vehicles while recording the crash scene details by the police. Further, the issue of hit-and-run crashes is prevalent in many parts of the world. For instance, in the USA, fatalities caused by hit-and-run crashes increased by 13.7% from 2009 to 2011 (NHTSA, 2012). India, which has a dubious distinction of leading the world in road traffic fatalities (MoRTH, 2018), also has a significant share of hit-and-run crashes, approximately 15% (69,822) in 2018. In terms of fatalities caused by hit-and-run (H&R) crashes, the trend is alarming (see Figure 1) and constituted around 19% (28,619) of total fatalities in 2018 (MoRTH, 2019). Further ahead, more than 22,000 persons suffered a grievous injury in hit-and-run crashes in 2017 alone (MoRTH, 2018), and a rising trend is also observed in recent years as per crash data (MoRTH, 2019). This is critical since approximately 35% of fatalities occur within 1-2 hours of crash occurrence (Roger P. Roess, 2004).

Road crashes ripple effects are evident in the nation's economy and well-being. About 3-5% of India's gross domestic product (GDP) is lost yearly because of road crashes (World Bank, 2020). Also, to achieve the global road safety target of a 50% reduction in fatalities, India will need an additional investment of USD 109 billion over the 2021-2030 decade (Bandyopadhyay et al., 2020). It is pertinent to point out the cost incurred concerning the hit-and-run crashes. As per MoRTH (2022) notification, in case of a hit-and-run crash, a compensation of Rs. 2 lacs is to be provided for death and Rs. 50,000 in case of grievously injured. Therefore, it is clear that hit-and-run crashes pose a tremendous economic burden to low and middle-income countries like India and necessitates an urgent response.

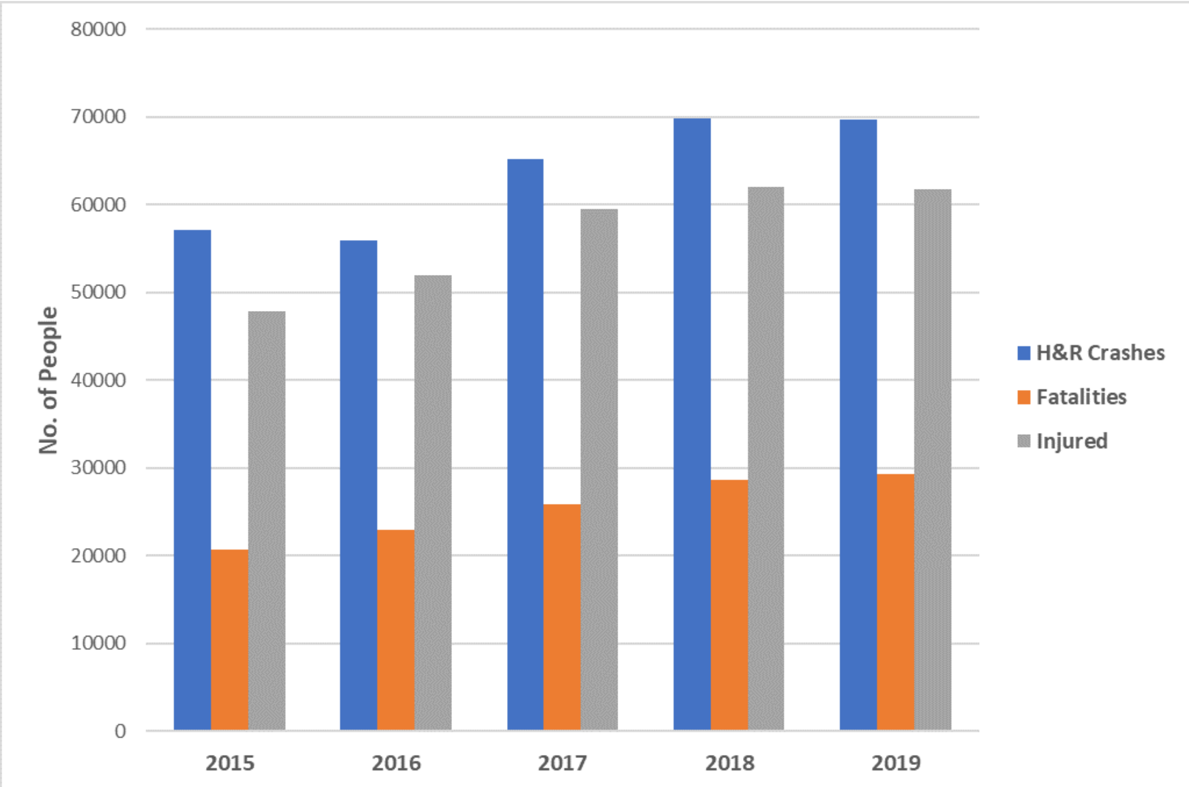

Fig. 1. Hit-and-run crashes scenario in recent years in India

The situation is aggravated further due to the unavailability of authentic or comprehensive crash data on offender vehicles in hit-and-run crashes in most jurisdictions. As a consequence, it becomes difficult to devise prevention strategies.

Based on the above discussion, it is evident that hit-and-run crashes pose a significant challenge and threat to the road safety situation in India. Therefore, the objective of this work is to identify the unknown striking vehicle type in a hit-and-run crash; it can be really helpful in developing strategic countermeasures to overcome this issue.

2. LITERATURE REVIEW

2.1 Hit-and-run crashes

Road traffic injuries and fatalities disproportionately impact low and middle-income countries (LMICs) like India (Dandona et al., 2020). To get insights into the underlying causal factors, the findings of a study commissioned by the World Bank in different states of India are important (World Bank, 2021). The study highlighted that (i) fatality post-crash is higher in low-income households than high-income households since most of them belong to the vulnerable road user group (pedestrians, MTWs, cyclists) and are involved in hit-and-run road crashes, (ii) low rates of insurance coverage (only one-third of the truck drivers in the study knew about third-party liability insurance) and in addition, lack of legal awareness among heavy vehicle (truck) drivers, due to this most truck drivers do not report the road crashes.

Hit-and-run crashes are a common scene in many countries around the world, including developed countries. Many past studies have explored the area in several ways. For instance, few studies have identified the causal factors in hit-and-run crashes (Tay et al., 2009; MacLeod et al., 2012; Zhang et al., 2014). Others have tried to understand the offender driver's decision to flee after the crash (Solnick and Hemenway, 1995; Tay et al., 2008; Kim et al., 2008). It is important to note that a reliable and comprehensive crash database is the primary requirement to perform these tasks. Low-and middle-income countries like India, where the crash reporting system is poor and which already suffer from having very few details (temporal, environment, vehicle, driver) about the road crash in the database. The possibility of achieving reliable results is meager and far-fetched.

The existence of limited studies originating from India despite the growing share of hit-and-run crashes is evidence of that. Recently, a study by Sivasankaran and Balasubramanian (2020), investigated the factors contributing to pedestrian hit-and-run crashes in India and found that seasonal (summer and winter), area type (urban area), and dark unlighted conditions increase the tendency of offender/ striking vehicle to flee from the road crash spot. However, the main issue in such hit-and-run crashes is that of not having knowledge of the striking vehicles, which are therefore reported as unknown vehicles in the crash data. Only a few studies (Jha et al., 2021) have recently explored this research area and have predicted missing information, such as unknown vehicles in a hit-and-run accident, using artificial intelligence-based models. However, the prediction accuracy of the models used was very less. The present study work is an attempt to build upon and extend the existing knowledge in predicting the unknown vehicles in hit-and-run crashes by employing other recent techniques to further facilitate in improving the road safety situation.

2.2 Prediction Models

Broadly, there are a total of four types of machine learning models that are used to perform the prediction analysis. They are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Kang and Jameson, 2018). In this study, as the output of the dataset (striking vehicle type) contains known and unknown striking vehicles so, the main aim of the present work is to predict the unknown striking vehicle type involved in the road crash using the available crash dataset. Hence, a supervised ML model is appropriate, but unsupervised models can also be applied to cluster the crash data and then apply any supervised ML model for the predictions.

Some of the supervised ML models which can be applied to the crash data are mentioned below.

1. Logistic regression: - It works like the Linear regression model, but the outcome of the linear regression model is continuous. However, in logistic regression, the output is categorical (Wright, 1995). This is a classification ML model used to predict the class of the unknown record. It can classify the instance into two or more two classes, also named a multiclass classification problem. It works based on Maximum likelihood estimation (Czepiel, 2002) and uses Log Loss as the Loss function to learn during training.
2. Linear discriminant analysis (LDA): - This is a supervised machine learning algorithm that is widely used in the dimensionality reduction of the data (Tang et al., 2005). The main aim of this algorithm is to reduce the within-class scatter, i.e., the similarity in the features for one class is high, and the other task is to increase the between-class scatter, which means the similarity in between the two classes is as low as possible (Izenman, 2008). It separates the data points of different classes and projects them on the perpendicular plane (e.g., if we are working on three-dimensional data, then projecting this data on a plane on which we find the maximum separation in the between-class scatter), leading to dimension reduction. And on top of it, we can apply any machine learning algorithms to the transformed data.
3. K-nearest neighbor (KNN): - It is also known as a lazy learning algorithm, which means it actually didn't learn anything and also didn't require any kind of training; it just calculates the euclidean distance from the given record with all the records present in the data (Cunningham and Delany, 2021). After finding the distances between all the records with the given one, The algorithm then selects the K-nearest data points, where K is a user-defined constant, and assigns the query point to the class that has the most representatives within the nearest neighbors (for example, car, car, truck, truck, truck in this K = 5 and returns 'car' as output for that instance) if its a classification problem.

   The KNN algorithm is also used in imputing the missing values, as done by Murti et al. (2019); this study examines the performance of an imputation method using the KNN algorithm to handle missing data. The results show that the accuracy of the imputed dataset is similar to that of a complete dataset.
4. Decision Tree: This is also known as CART (classification and regression technique), i.e., used for both regression and classification tasks (Crawford, 1989). The basic intuition behind this algorithm is that it splits the decision in terms of 'yes' or 'no' and divides the data into subsets; this process is done on every node, and the leaf node of the tree is the outcome that we are looking for. It uses the Gini index and entropy to select the splitting value at every node (Charbuty and Abdulazeez, 2021). Gini is the measure of the impurity of the data at that node, and entropy is the measure of the variability of the data at that node. Hence, if we are selecting Gini, it must be the minimum, and if we are using the entropy, then the difference between the entropy before and after splitting has to be maximum.
5. Support vector machine (SVM): - It is also a supervised machine learning algorithm whose main aim is to draw a hyperplane in between the two classes of the given data (Wang, 2005) so that it acts as the decision boundary for the upcoming data whether it lies in which side of the hyperplane. This can also be used in regression tasks, regression tasks include building a residual insensitive tube for regressing the outcomes, but here in the present study, we need to classify the categories. It uses several kernels which transform the data to a higher dimension as needed for building the hyperplane. Some

of them are ‘linear,’ ‘rbf’ radial basis function (used to increase the dimension of the data), ‘polynomial’ (Suthaharan, 2016), etc.

6. Naive Bayes (NB):- Naive Bayes is a machine learning algorithm for classification problems, which is based on the Bayes theorem. It states that the probability of an event occurring is equal to the product of the probability of the event given some evidence and the prior probability of the event (Zhang, 2004). Naive Bayes makes use of this theorem to classify data into different categories. It assumes that all features are independent of each other, which makes it a simple and fast algorithm. Naive Bayes has been used in many applications, such as spam filtering, text classification, and medical diagnosis. It is also widely used in natural languages processing tasks such as sentiment analysis and document categorization.
7. Random Forest:- Random forest is a powerful machine-learning model that is used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to create a more accurate and robust model. The random forest algorithm works by randomly selecting a subset of features from the training dataset and then building multiple decision trees using those features. Each tree is then used to make predictions on the test data, and the results are combined to form a single prediction (Breiman, 2001). Random forests are known for their accuracy, robustness, and scalability, making them a popular choice for many machine-learning tasks.

In the past, various machine learning models such as logistic regression, support vector classifier, KNN, Naïve Bayes, and decision trees have been used for different purposes. For example, (Rezapour et al., 2020) employed logistic regression and a decision tree to analyze the injury severity of motorized two-wheeler (MTW) at-fault crashes. (Jamal et al., 2021) compared the eXtreme Gradient Boosting (XGBoost) model to traditional machine learning algorithms for crash injury severity analysis using data from 13,546 motor vehicle collisions in Riyadh, KSA. Results indicated that XGBoost outperformed other models in terms of predictive performance and individual class accuracies. Several studies (Iranitalab and Khattak, 2017), (Zhang et al., 2018), (Wahab and Jiang, 2019), (Komol et al., 2021) also conducted similar studies to predict crash severity using statistical and machine learning methods for MTWs and vulnerable road users, respectively. These studies have demonstrated the potential of machine learning models when applied to crash data.

As per the authors’ best knowledge, only one study has tried to predict the unknown striking vehicle type in hit-and-run cases (Jha et al., 2021). The authors compared the above models to predict the unknown striking vehicles in hit-and-run cases. Based on how well it worked in their case, the Support vector classifier is the best because it works best with space data. Therefore, the present study attempts to extend the existing literature by using robust machine learning models and to bridge the gap by identifying the unknown striking vehicle type for hit-and-run crashes involving one of the vulnerable road users (MTWs).

3. METHODOLOGY

The main aim of this study is to predict the striking vehicle type in the hit-and-run crash, which is reported as ‘unknown’ in the crash data. This is done by using several classification machine learning algorithms, as seen in Figure 2. Here the classification problem is not inclined towards either of the situations like we can bear a false negative (e.g., cancer patient prediction) or a false positive (e.g., criminal prediction); hence the present study is more towards the accuracy of the model, not towards the recall and precision.

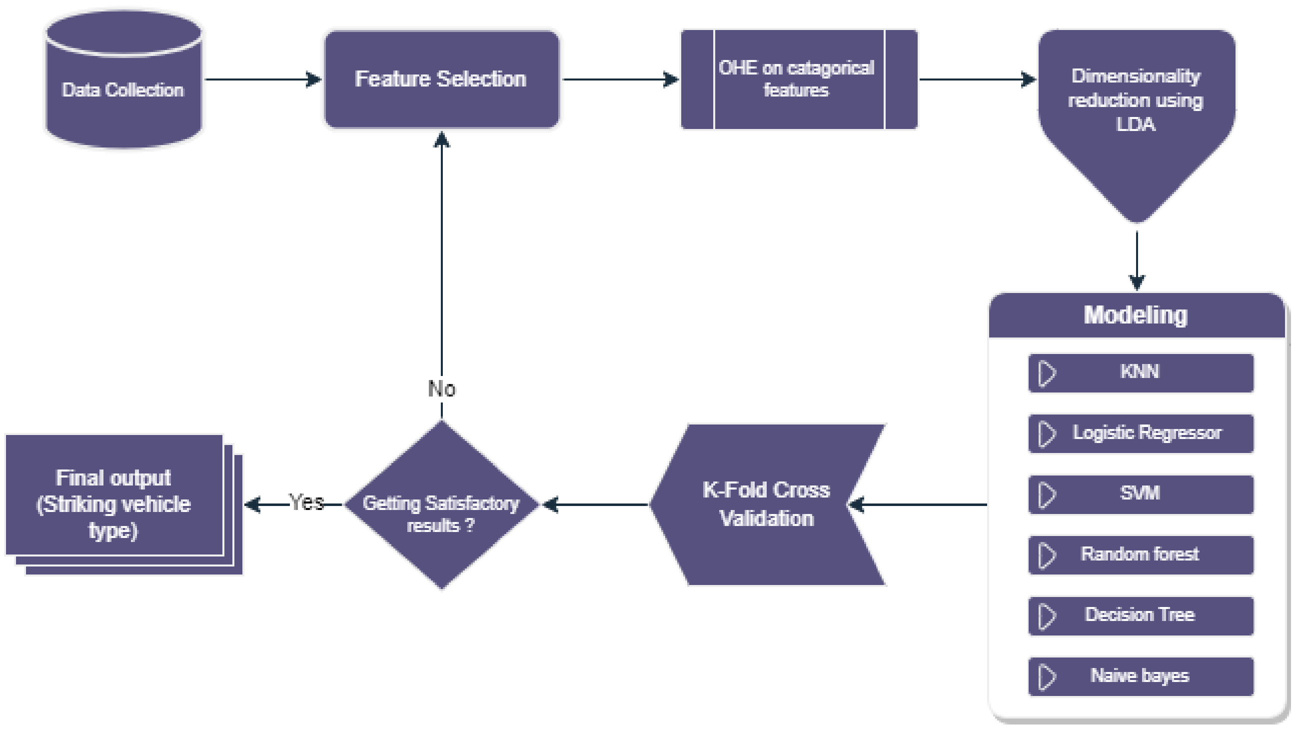

Fig. 2. Methodology Flow for the Study

The steps involved in the development of the model are as follows-

- At first, a feature selection process, i.e., selecting the most reasonable features responsible for output, is performed. Initial data had a set of 12 features; they are MTW Crash Severity, Day, Time of the crash, Season, Road type by geometry, Road location by type of neighborhood, Median Presence, Collision type, MTW by Engine Capacity, Pillion Passenger Presence, MTW Rider Gender, and MTW Rider Age. Out of 327 instances, 120 are unknown, i.e., a prediction model is required for these instances. Hence a total of 207 cases are left for training and testing the proposed model.
- After doing the feature selection process, out of 12 features, 11 are categorical features, which can not be used directly while building the model. Linear discriminant analysis is the best-suited technique for dimension reduction if we are dealing with the categorical output variable, and finally, the model building and validation part is carried out, as seen in Figure 2.
- On these features, we have applied one hot encoding, which converts these 11 features into 38 feature spaces, and we have a total of 207 instances for model building. One hot encoding makes the data very sparse, and predictions on the sparse data set are not easier for many machine learning algorithms (except SVM); hence we need to reduce the dimension of the dataset.
- For dimensionality reduction, we applied LDA, which is very useful when we have categorical outputs. It is a supervised algorithm that will help us in further model building for prediction. After applying LDA, the feature map is reduced to six, and cases are the same as before, i.e., 207.
- After the LDA, we applied six ML models: Decision tree, SVC, Random Forest classifier, Naive Bayes, KNN, and Logistic regression. These models are being compared with the help of a cross-validation algorithm.

- Finally, the validation of these machine learning models is carried out using 10-fold, 5-fold, and 4-fold cross-validation, as the crash data set is limited, and seven classes are to be predicted, so it is better to check the model accuracy using various cross-validations.

3.1 Dimensionality reduction using linear discriminant analysis (LDA)

The curse of dimensionality is a phenomenon that occurs when the number of dimensions in a dataset increases, leading to an exponential increase in the amount of data needed to represent the data accurately. This phenomenon has been studied extensively in the research literature by Bellman and Kalaba (1959); this paper showed that as the number of dimensions increases, it becomes increasingly difficult to accurately represent the data due to the large amount of data needed. Furthermore, they have demonstrated that certain techniques, such as principal component analysis (PCA), can be used to reduce the dimensionality of a dataset and thus reduce the amount of data needed for accurate representation (Bellman and Kalaba, 1959).

Dimensionality reduction is a process of reducing the number of features or variables in a dataset while preserving the most important information. It is an important step in data pre-processing and can be used to reduce the complexity of a dataset, improve the accuracy of machine learning models, and reduce the time required for training (Van Der Maaten et al., 2009). Linear Discriminant Analysis (LDA) is one of the most popular techniques for dimensionality reduction if we are working with a classification problem.

3.2 Cross Validation

After reducing the dimension of the crash data, several machine learning models are developed and compared to their generalization by cross-validation. If we are dealing with biased data, then it is difficult for a model to perform well on testing data. A good model must not overtrain on training sample because it may lead to overfitting. Hence, we need a generalized model. A K-fold cross-validation approach is applied to all the modes while learning from training and finding the test set. Cross-validation is a technique that is used to validate the built model, whether it is generalized or not (Browne, 2000). This is done by splitting the dataset into several folds, and we used one fold at a time for testing and all the remaining one for training. This process tests the model to determine whether it performs well on these several operations. For example, if we are talking about the 10-fold cross-validation, then it means that we have folded the crash data in 10 folds, and out of them, 9 are used for training, and one is for testing. This process is done 10 times because we have 10 folds, i.e., every fold is used as testing data once.

The ‘K’ in K-fold cross-validation stands for the number of folds or partitions that the data is divided into (Anguita et al., 2012). K is typically an integer value greater than 2. In the present study, we have used 10-Fold, 5-Fold, and 4-fold cross-validation algorithms. Among these algorithms, 5-fold has varied advantages over others. For instance-

- 5-fold cross-validation is less prone to overfitting than 10-fold cross-validation since it uses a smaller portion of the data for training and testing.
- 5-fold cross-validation can provide more accurate results than 10-fold cross-validation since it uses a larger portion of the data for training and testing by ensuring that each fold contains an equal representation of all classes in the dataset; hence the biasedness of the model is reduced.

Further ahead, accuracies for all the operation is noted, and the overall mean of these accuracies can be treated as the accuracy of the model, and we can also get the approximate standard deviation of the accuracy from the noted accuracies (Maglogiannis, 2007). Accuracy is calculated with the help of a formula for the categorical output, as mentioned below. This process is useful to reduce bias and variance in the model; in other words, overfitting and underfitting are addressed by this validation test. Here, true positive is denoted as TP, true negative as TN, false positive as FP, and false negative as FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

4. STUDY AREA AND ROAD CRASH DATA

4.1 Study area

The capital and megacity of India, i.e., Delhi, was selected based on the crash trends in recent years. Among the million-plus cities (in terms of population), Delhi leads in road traffic fatalities (MoRTH, 2015, 2016, 2017, 2018, 2019). Further, the trend in hit-and-run road crashes in Delhi in recent years is also worrisome (see Figure 3), being a highly urbanized city.

The safety record of MTWs, which dominate traffic streams in Delhi, with more than 60% share (DTP, 2018), is also a concern. As per crash statistics, MTW users were victims in 1 of every 3 deaths or injuries (DTP, 2018). Further, Delhi traffic police also practice a rationale approach wherein they identify the accident-prone zones every year for each vulnerable road user (pedestrian, MTWs, cyclists) based on the following criterion: (i) 3 or more fatal crashes within the circle of diameter of 500 meters or (ii) 10 or more total crashes in the same region or location.

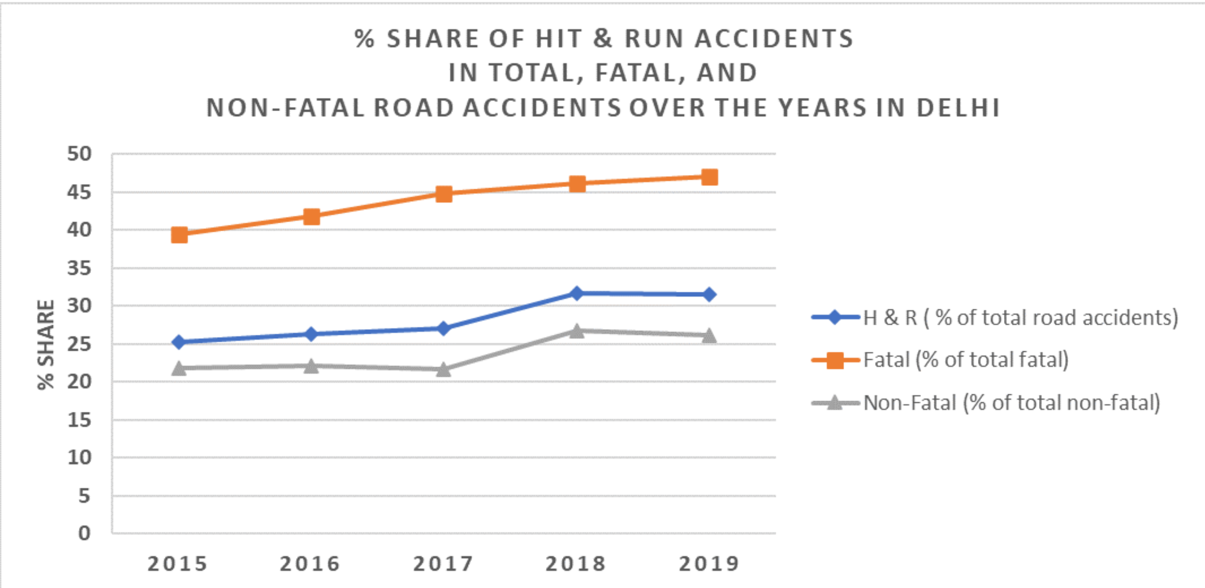

Fig. 3. Hit-and-run crashes scenario in the study area Delhi

For the present study, based on the pre-defined criteria, they have identified the crash-prone zones for the MTWs for a period of 3 years (i.e., 2016 - 2018), and from which a total of 327 crash first information reports (FIRs) from the MTW accident-prone zones are retrieved and examined in this study.

4.2 Road Crash Data Description

Delhi traffic police record the road crash data in the first information reports (FIRs). Crash FIR data fields provide information about the crash date, day, time, location, a brief description of the crash, and so on. The inputs for data fields are obtained from the police investigating officer, crash victim, offender vehicle driver, pillion passenger, if any, with the victim, or eyewitness. From the road crash FIRs for a study period (2016-2018), the following variables are retrieved from 327 road crash FIRs involving MTWs:

a. Temporal information: Month, day, and time of road accident
b. Roadway information: Type of road geometry, type of neighbourhood, median presence
c. Crash-specific information: Striking vehicle type, collision type, hit-and-run status, the severity of crash (fatal, non-fatal)
d. Road user information: Victim (MTW rider) gender, age, the pillion passenger presence

5. RESULTS AND DISCUSSION

5.1 Road crash pattern in MTW accident-prone zones

Road Crash Severity- From 2016 to 2018, there were 327 crashes in MTW accident-prone zones in Delhi, which include 111 (33.94%) fatal and 216 (66.06%) non-fatal MTW crashes. It is evident from Figure 4 that a constant trend exists in MTWs fatalities. In terms of hit-and-run crashes, the year 2016 (56, 42%) had the maximum number of hit-and-run crashes involving MTWs. Further, 21 (55%) of fatalities for the year 2016 were reported in hit-and-run crashes. This shows the menace of hit-and-run crashes in the case of vulnerable road users like MTWs.

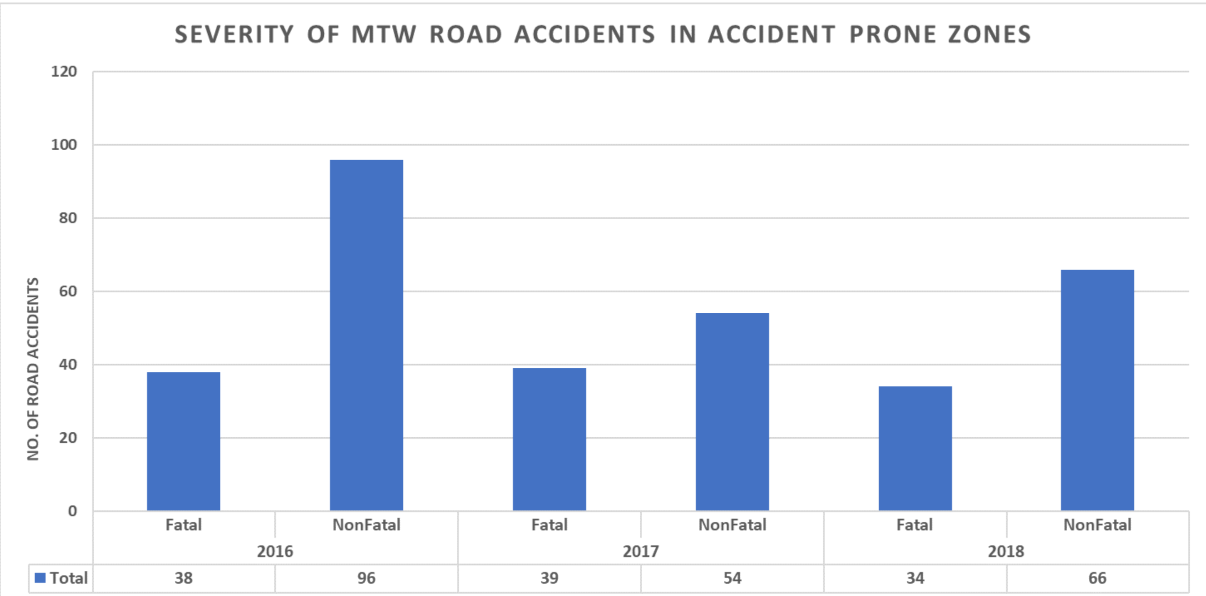

Fig. 4. Severity of crashes in MTW Crash Prone Zones

Temporal Trend in MTW crashes- Figure 5 shows the number of MTW crashes by the time period in a day. It is evident that the number of MTW crashes peaked during the night hours (9 pm-12 am) and was lowest during the post-midnight (3 am to 6 am) and early morning (6 am to 9 am) hours when the level of MTW traffic is likely to be lower. In terms of hit-and-run crashes, night-time is dangerous for MTWs, especially from (9 pm-12 am) and (12 am-3 am), and constituted about 61 MTW crashes, i.e., 50% of total hit-and-run crashes. Moreover, the data shows that most hit-and-run crashes involving MTWs at night were fatal.

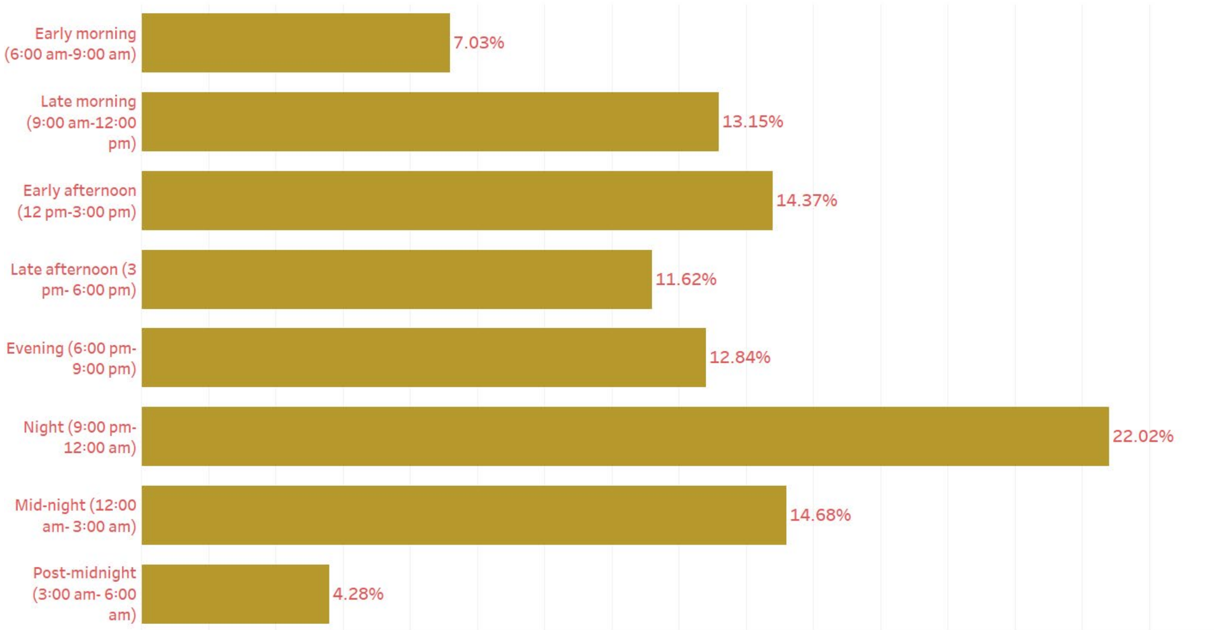

Fig. 5. Distribution of MTW hit-and-run crashes with respect to time

Spatial Trend in MTW crashes- Figure 6 shows the distribution of MTW crashes based on location in urban areas. Based on crash location, 128 (39.14%) of total MTW crashes occurred on flyovers which are the most prone locations for MTW crashes. Urban mid-blocks were the second most prone location for MTW crashes, with 94 (28.75%), followed by signalized

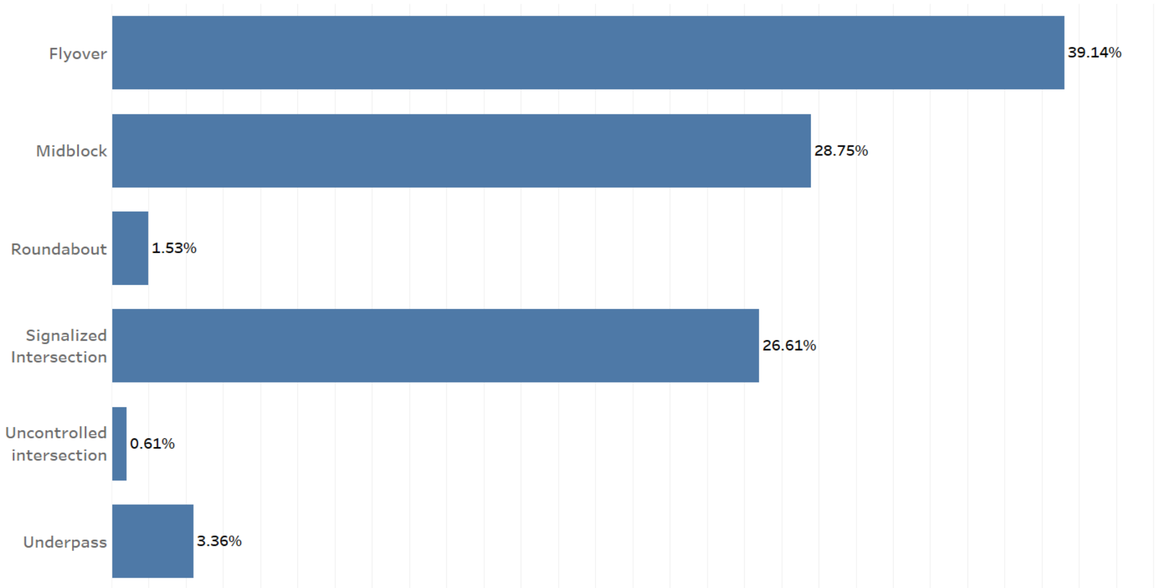

Fig. 6. Spatial distribution of MTW road crashes

intersections with 87 (26.61%) MTW crashes. In terms of hit-and-run crashes, where the striking vehicle is unknown, flyovers (54, 45%) dominate, followed by urban midblock (36, 30%) and signalized intersections (23, 19.2%), respectively.

Striking Vehicles in MTW Road crashes- Figure 7 shows the distribution of MTW crashes based on the striking or impacting vehicle. Based on the type of striking vehicle in MTW crashes, cars are the most reported and accounted for 85 (26%), followed by the truck with 36 (11%). A significant proportion of LMV (29, 9%) was also involved as the striking/ impacting vehicle. Most importantly, hit-and-run crashes (120, 36.7%) dominate the MTW crashes; these are the crashes in which the striking vehicle is unknown.

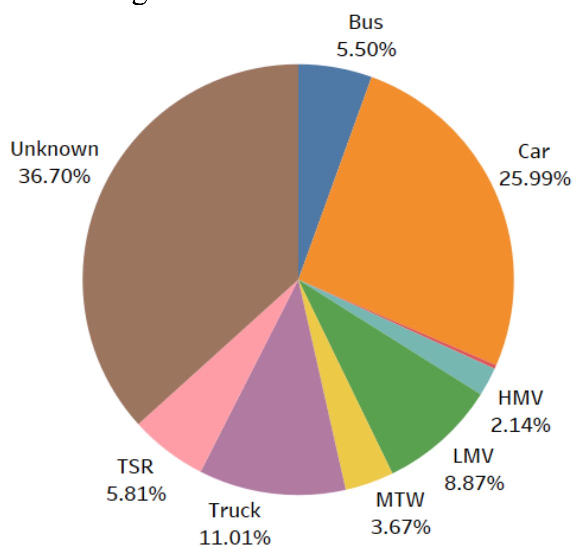

Fig. 7. Share of striking vehicle by type in MTW road crashes

5.2 Selection of Supervised Classification Model

The results after building several machine-learning models can be seen in Table 1. The cross-validation of these models is carried out using 10-fold, 5-fold, 4-fold cross-validation while doing the validation using 10-fold, i.e., 90% of data is used for training, and 10% for testing is giving us overfitted results because here we have a small dataset with 7 classes to predict so while splitting the data in such a ration may lead to biased training sample. Hence, a balanced splitting is necessary, so using 5-fold validation, i.e., 80-20 splitting of data, addresses the overfitting and also reduces the deviation of accuracy in all the operations carried out during the cross-validation.

Based on the prediction accuracy and standard deviation, we can infer that the decision tree has poor results as compared to the remaining models because, in most cases, decision trees overfit on the training set and lead to poor performance on the testing data, so an ensemble technique (i.e., a random forest) which is a combination of several decision trees; it is always a better option when compared with the decision tree because the predictions from multiple decision trees, i.e., multiple machine learning models and getting a combined outcome of all of them leads to a generalization of the model. Further, Naive Bayes only uses past events to predict the future, and there can be the case when previously some of the events never occurred; hence this algorithm gives so much accuracy deviation. And in the case of KNN, it does not give importance to any feature; it simply calculates the distance between the instances and returns the nearest one and not giving better results; hence it is also not able to generate a good relationship with the dependent and independent variables.

The remaining models show accuracy within a range of 51-56%, with the highest being of Logistic regressor (55.76%) and next random forest (54.55%), but it can also be noted that the standard deviation in the accuracies while doing cross-validation is minimum in the random forest which states that this model is giving us consistent results in all the validation operation, i.e., we can rely on these outcomes as compared with the other. Therefore, based on the accuracy and standard deviation of the models, the Random forest model can be selected as the best-fit model because it has the least standard deviation in accuracy, which means it is the most consistent model among all. Hence we are selecting the Random forest for further predictions.

Table 1. Cross validation with 5 Folds : 80% for training and 20% for testing

S.No.	Model	Prediction Accuracy (%)	Standard Deviation
1	Decision Tree	41.21	7.32
2	SVC	52.73	4.54
3	Naive-Bayes	51.52	9.19
4	Random Forest	54.55	2.71
5	Logistic Regression	55.76	4.11
6	KNN	41.82	3.53

5.3 Predicting Unknown Striking Vehicle Type in the Hit-and-run Crashes

Figure 8 shows the predicted results; clearly, it exhibits that in the hit-and-run crashes (120, 36.7%) involving MTWs as the victim, car drivers had the major share as striking/ offending vehicles (55, 46%) followed by trucks (38, 32%).

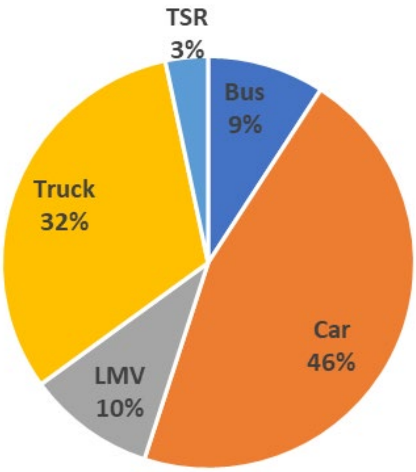

Fig. 8. Predicted Striking vehicle type

For hit-and-run road crashes based on the time period of the day (see Figure 9), at night time (9 pm-12 am) which was found most dangerous, car drivers had the major share, followed by truck drivers. The trend was the opposite during the midnight (12 am- 3 am) period, where the truck drivers had a major share in hit-and-run crashes involving MTWs as the victim. These findings necessitate urgent tactical decisions (enforcement, education, medical care) based on the critical time period identified for hit-and-run crashes involving MTWs.

Similarly, for hit-and-run crashes based on urban location (see Figure 10), flyovers which had the maximum hit and- run crashes involving MTWs, car drivers had the major share in hit-and-run crashes, followed by truck and light motor vehicles (LMV). On midblock, surprisingly, truck drivers had the major proportion in hit-and-run crashes, followed by cars and buses. This is critical information and necessitates enforcement as well as engineering intervention. On signalized intersections, cars and trucks were the prime offending/ striking vehicle in hit-and-run crashes involving MTWs as the victim. Overall, it was found that car and truck drivers had the major share in hit-and-run crashes involving MTWs therefore, enforcement drives can be planned accordingly.

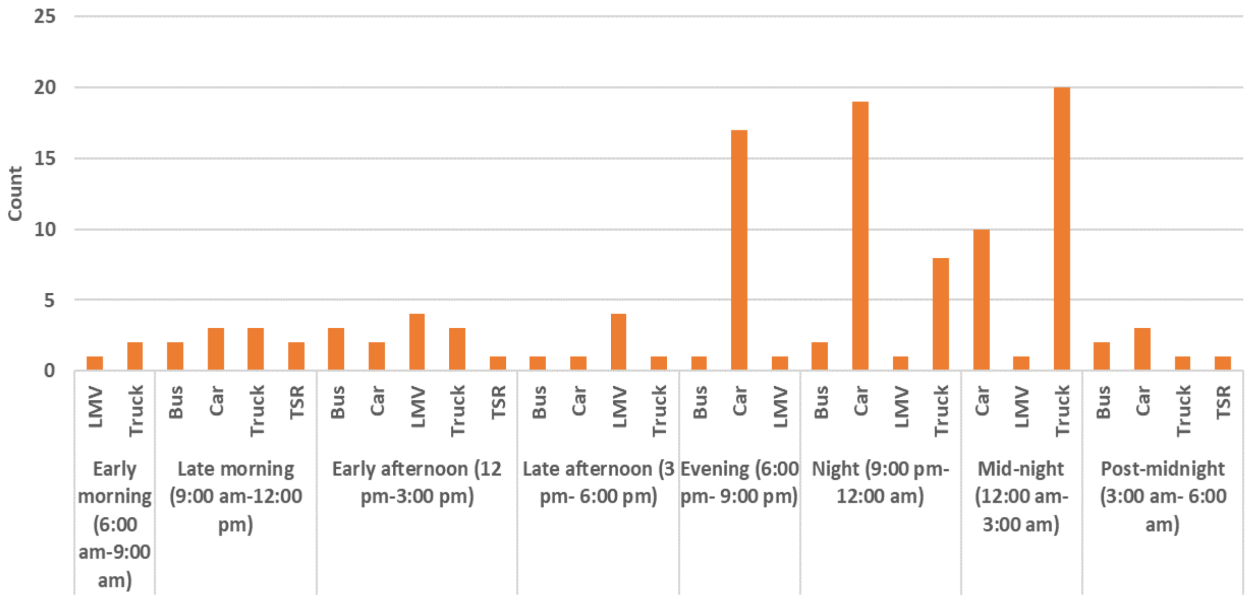

Fig. 9. Predicted striking vehicle type based on time interval

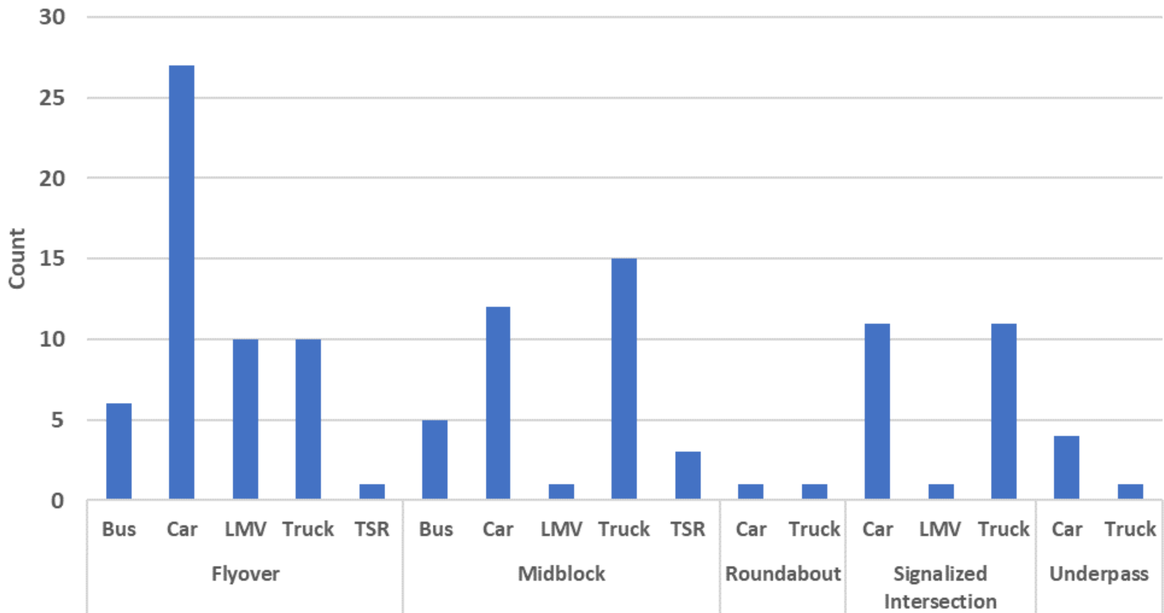

Fig. 10. Predicted striking vehicle type based on urban locations

5.4 Discussion

If we look at the maximum accuracy of the machine learning models built in (Jha et al., 2021), we see the following: CART (decision tree) got 26% in Amritsar on test data, but we got 41.21% in Delhi, which is much better. This shows that there is a high chance of overfitting in their case, since they only had 263 training and testing samples to build and tune the model, and the decision tree is very prone to overfitting if it is not properly tuned due to the high dimensionality of the data. Then, in every other situation, the support vector machine got the highest accuracy, which was 45% in Ludhiana, 38% in Bhopal, 37% in Vizag, and 44% in Agra, as done by Jha et al. (2021). In the present study, using the support vector classifier, we got 52.73% accuracy in Delhi, which was better than all of the above. In addition to their machine learning models, we've also made a random forest classifier, which has the benefits we've already talked about. So, by using this model, we got an accuracy of 54.55%, which was higher than that of the support vector classifier and gave consistent results when cross-validated. So, the best way to make predictions is to use a random forest, as shown in the present work.

Grade-separated intersections (flyovers) had the maximum share in MTW crashes in Delhi as per crash data. In this respect, the study by (Gupta et al., 2010) provides interesting insights based on the comparison of the mean speed of vehicles post-construction of the AIIMS flyover in Delhi. They found that the speed of vehicles increased by 21.5%, 22.6%, 15%, and 31.6%, respectively, for heavy vehicles, cars, three-wheelers, and motorized two-wheelers, respectively. This underlines the fact that vehicles, including MTWs, tend to overspeed on grade-separated intersections (flyovers), which increases not only the chance of crashes but also the severity since they interact with large vehicles. Another important point is that typically, heavy vehicles are allowed in the night, and they are assumed to be loaded; thus, the driving maneuver of heavy vehicles is different at up/downgrades of the flyover; since this information is missing from the crash data, it should be looked in the future studies.

Further, the spatial trend of MTW crashes revealed that midblock is the second most accident-prone location for MTWs, followed by signalized intersections. On urban midblock and intersections, there is a widespread belief that MTWs are more difficult to detect in traffic than any other motorized vehicle due to conspicuity issues. Earlier studies (Haque et al., 2009; Hurt et al., 1981; Mannering and Grodsky, 1995) of individual collisions involving MTWs, had indicated that drivers who violate MTW right-of-way often claim not to have seen them before the collision ("looked but failed to see"). In this regard, (Tiwari et al., 1998) performed conflict analysis for the prediction of fatal crash locations in mixed traffic streams and suggested segregation and traffic calming techniques development with special reference to motorized two-wheelers.

Similarly, special treatment at intersections is given to MTWs in some parts of the world to facilitate their clearance from the intersection quickly and reduce delays to other vehicles. In Taiwan, motorcycles are allowed to store behind the stop line at a few intersections (Lee, 2008). In Chennai, India, the study by Asaithambi et al. (2015) suggested that for MTW-dominated traffic (70% MTWs) at signalized intersections, the discharge rates can be inherently increased (less delays) with the provision of exclusive stopping space for motorized (ESSM) two-wheelers near the stop line.

6. CONCLUSION AND RECOMMENDATIONS

Road crashes are endangering the lives of millions worldwide, especially road users in low and middle-income countries like India. The socio-economic cost of road crashes is also immense; unfortunately, vulnerable road users like MTWs, etc., bear the brunt of this. Another neglected issue is that of hit-and-run crashes, wherein no accountability can be fixed between participating vehicles since the information on the offender's vehicle or a striking vehicle is unknown in the crash data.

The present study focuses on identifying the unknown striking vehicle type in hit-and-run crashes involving MTWs so that prevention strategies can be developed accordingly. The work is carried out by first identifying the most important features from the road crash dataset, followed by the dimensionality reduction of the data. After this, different supervised learning models (logistic regression, support vector classifier, KNN, decision tree, Naive-Bayes, random forest) are applied for the prediction of the striking vehicles. The validation of these models was carried out using the K-fold cross-validation algorithm. In this study, the ensemble machine learning (Random forest) model best predicted the unknown striking vehicle type, among other models.

Based on the prediction of the striking vehicle type in hit-and-run crashes involving MTWs as the victim, it was found that car and truck drivers had a major share in the hit-and-run crashes. Further, for hit-and-run crashes based on the time period of the day, night time (9 pm-12 am) was found most dangerous. The model predicted car drivers as the striking/offender vehicle in a significant proportion of night-time hit-and-run crashes, followed by truck drivers. The trend was the opposite during the midnight (12 am- 3 am) period, where the truck drivers had a major share in hit-and-run crashes. Further ahead, for hit-and-run crashes based on urban locations, flyovers had the maximum number of hit-and-run crashes involving MTWs as the victim. It was found that car drivers had the major share in hit-and-run crashes on flyovers, followed by truck and light motor vehicles (LMV). On midblock, interestingly, truck drivers had the major proportion in hit-and-run crashes, followed by cars and buses. This necessitates for segregation of MTWs from heavy vehicle traffic on urban roads. Further on, signalized intersections, cars, and trucks dominated as the offending/striking vehicle in the hit-and-run crashes involving MTWs as the victim.

Recommendations based on study findings:

- Special enforcement drives should be conducted during night-time, especially from 9 pm- 3 am in MTW accident-prone zones
- Additional allocation of medical ambulances in the MTW accident-prone zones during the critical night-time for necessary post-crash care to victims
- Training and awareness programs for car and heavy vehicle drivers emphasizing responsible driving during night-time and the importance of golden hour in case of road crashes and how it can reduce the fatality risk
- Sensitization of road users about the good samaritan laws and how it protects them when reporting about the road crash victim
- Use of bright/reflective clothing for MTW users, especially during the night, to improve their visibility to other road users
- Reduction of posted speed limits on grade-separated intersections (flyovers) for motorized vehicles to reduce the severity of road crashes for MTWs

ACKNOWLEDGMENTS

The authors would like to convey thanks to the Accident Research Cell, Delhi Police, Traffic Police (HQ), Todapur, Delhi, for sharing road crash data used in this research work. The opinions, findings, and conclusions expressed here are those of the authors.

REFERENCES

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., (2012). The ‘k’ in k-fold cross validation, in: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), i6doc. com publ. pp. 441–446.

Asaithambi, G., Kumar, R., Sivanandan, R., (2015). Microscopic simulation for modeling exclusive stopping space for motorcycles under non-lane based mixed traffic conditions. European Transport TrasportiEuropei57.URL:http://www.istiee.unict.it/europeantransport/papers/N57/P05_57_05_2015.pdf.

Bandyopadhyay, A., Nora, J.E., Bose, D., Srinivasan, K., Woodrooffe, J.H.F., Surie, N., Bliss, A.G., (2020). Delivering Road Safety in India: Leadership Priorities and Initiatives to 2030 (English). TechnicalReport.TheWorldBank.URL:http://documents.worldbank.org/curated/en/827641581916488024/Delivering-Road-Safety-in-India-Leadership-Priorities-and-Initiatives-to-2030.

Bellman, R., Kalaba, R., (1959). A mathematical theory of adaptive control processes. Proceedings of the National Academy of Sciences 45,1288–1290.

Breiman, L., (2001). Random forests. Machine learning 45, 5–32.

Browne, M.W., (2000). Cross-validation methods. Journal of mathematical psychology 44, 108–132.

Charbuty, B., Abdulazeez, A., (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends 2, 20–28. doi:10.38094/jastt20165.

Crawford, S.L., (1989). Extensions to the cart algorithm. International Journal of Man-Machine Studies 31, 197–217. doi:10.1016/0020-7373(89)90027-8.

Cunningham, P., Delany, S.J., (2021). K-nearest neighbour classifiers-a tutorial. ACM Computing Surveys (CSUR) 54, 1–25. doi:10.1145/3459665.

Czepiel, S.A., (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Technical Report. URL:https://czep.net/stat/mlelr.pdf.

Dandona, R., Kumar, G.A., Gururaj, G., James, S., Chakma, J.K., Thakur, J.S., Srivastava, A., Kumaresh, G., Glenn, S.D., Gupta, G., Krishnankutty, R.P., Malhotra, R., Mountjoy-Venning,W.C., Mutreja, P., Pandey, A., Shukla, D.K., Varghese, C.M., Yadav, G., Reddy, K.S., Swaminathan, S., Bekedam, H.J., Vos, T., Naghavi, M., Murray, C.J.L., Dhaliwal, R.S., Dandona, L., (2020). Mortality due to road injuries in the states of india: the global burden of disease study 1990–2017. The Lancet Public Health 5, e86–e98. doi:10.1016/s2468-2667(19)30246-4.

DTP, (2018). Road Accidents In Delhi-2018. Technical Report. Delhi Traffic Police. URL: https://delhitrafficpolice.nic.in/road-accidents-delhi-2018.

Gupta, U., Chatterjee, N., Tiwari, G., Fazio, J., (2010). Case study of pedestrian risk behavior and survival analysis. Journal of the Eastern Asia Society for Transportation Studies 8, 2123–2139. doi:10.11175/easts.8.2123.

Haque, M.M., Chin, H.C., Huang, H., (2009). Modeling fault among motorcyclists involved in crashes. Accident Analysis & Prevention 41, 327–335. doi:10.1016/j.aap.2008.12.010.

Hurt, H.H., Ouellet, J., Thom, D.R., et al., (1981). Motorcycle accident cause factors and identification of countermeasures. Volume 1: technical report. Technical Report. United States. National Highway Traffic Safety Administration.

IHME, (2017). Global burden of diseases. Technical Report. The Institute for Health Metrics and Evaluation. URL: https://vizhub.healthdata.org/gbd-compare/.

Iranitalab, A., Khattak, A., (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention 108, 27–36. doi:10.1016/j.aap.2017.08.008.

Izenman, A.J., (2008). Linear discriminant analysis, in: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer New York. pp. 237–280. doi:10.1007/978-0-387-78189-1_8.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D., Ahmad, M., (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. International journal of injury control and safety promotion 28, 408–427.

Jha, A.N., Chatterjee, N., Tiwari, G., (2021). A performance analysis of prediction techniques for impacting vehicles in hit-and-run road accidents. Accident Analysis & Prevention 157, 106164. doi:10.1016/j.aap.2021.106164.

Kang, M., Jameson, N.J., (2018). Machine learning: fundamentals. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things , 85–109.

Kim, K., Pant, P., Yamashita, E.Y., (2008). Hit-and-run crashes. Transportation Research Record: Journal of the Transportation Research Board 2083, 114–121. URL: https://doi.org/10.3141/2083-13, doi:10.3141/2083-13.

Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M., Rakotonirainy, A., (2021). Crash severity analysis of vulnerable road users using machine learning. PLOS ONE 16, e0255828. doi:10.1371/journal.pone.0255828.

Lee, T.C., (2008). An Agent-Based Model to Simulate Motorcycle Behaviour in Mixed Traffic. Phd thesis. Imperial College London, UK.

MacLeod, K.E., Griswold, J.B., Arnold, L.S., Ragland, D.R., (2012). Factors associated with hit-and-run pedestrian fatalities and driver identification. Accident Analysis & Prevention 45, 366–372. doi:10.1016/j.aap.2011.08.001.

Maglogiannis, I.G., (2007). Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in ehealth, hci, information retrieval and pervasive technologies. volume 160. Ios Press.

Mannering, F.L., Grodsky, L.L., (1995). Statistical analysis of motorcyclists' perceived accident risk. Accident Analysis & Prevention 27, 21–31. doi:10.1016/0001-4575(94)00041-J.

MoRTH, (2015). Road accidents in India 2015. Technical Report. Ministry of Road Transport & Highways.

MoRTH, (2016). Road accidents in India 2016. Technical Report. Ministry of Road Transport & Highways.

MoRTH, (2017). Road accidents in India 2017. Technical Report. Ministry of Road Transport & Highways.

MoRTH, (2018). Road accidents in India 2018. Technical Report. Ministry of Road Transport & Highways.

MoRTH, (2019). Road accidents in India 2019. Technical Report. Ministry of Road Transport & Highways.

MoRTH, (2022). Notification issued for compensation of victims of Hit and Run motor accidents. Technical Report. Ministry of Road Transport & Highways. URL: **https://pib.gov.in/PressReleasePage.aspx?PRID=1801656**.

Murti, D.M.P., Pujianto, U., Wibawa, A.P., Akbar, M.I., (2019). K-nearest neighbor (k-nn) based missing data imputation, in: 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 83–88. doi:10.1109/ICSITech46713.2019.8987530.

NHTSA, (2012). Fatality Analysis Reporting System (FARS). Technical Report. National Highway Traffic Safety Administration. URL: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars/.

Rezapour, M., Molan, A.M., Ksaibati, K., (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. International Journal of Transportation Science and Technology 9, 89–99. doi:10.1016/j.ijtst.2019.10.002.

Roger P. Roess, Elena S. Prassas, W.R.M., (2004). Traffic Engineering. 3rd edition ed., Pearson Education International.

Sivasankaran, S.K., Balasubramanian, V., (2020). Investigation of factors contributing to pedestrian hit-and-run crashes in India. Journal of Transportation Safety & Security 14, 382–403. doi:10.1080/19439962.2020.1781313.

Solnick, S.J., Hemenway, D., (1995). The hit-and-run in fatal pedestrian accidents: Victims, circumstances and drivers. Accident Analysis & Prevention 27, 643–649. doi:10.1016/0001-4575(95)00012-o.

Suthaharan, S., (2016). Support vector machine, in: Machine learning models and algorithms for big data classification. Springer, pp. 207–235.

Tang, E.K., Suganthan, P.N., Yao, X., Qin, A.K., (2005). Linear dimensionality reduction using relevance weighted lda. Pattern recognition 38, 485–493. doi:10.1016/j.patcog.2004.09.005.

Tay, R., Barua, U., Kattan, L., (2009). Factors contributing to hit-and-run in fatal crashes. Accident Analysis & Prevention 41, 227–233.doi:10.1016/j.aap.2008.11.002.

Tay, R., Rifaat, S.M., Chin, H.C., (2008). A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes. Accident Analysis & Prevention 40, 1330–1336. doi:10.1016/j.aap.2008.02.003.

Tiwari, G., Mohan, D., Fazio, J., (1998). Conflict analysis for prediction of fatal crash locations in mixed traffic streams. Accident Analysis & Prevention 30, 207–215. doi:10.1016/S0001-4575(97)00082-1.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al., (2009). Dimensionality reduction: a comparative. J Mach Learn Res 10, 13.

Wahab, L., Jiang, H., (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLOS ONE 14, e0214966. doi:10.1371/journal.pone.0214966.

Wang, L., (2005). Support vector machines: theory and applications. volume 177. Springer Science & Business Media. doi:10.1007/b95439.

WHO, (2018). Global status report on road safety 2018. Technical Report. World Health Organization. URL: https://www.who.int/publications/i/item/9789241565684

World Bank, (2020). Guide for Road Safety Opportunities and Challenges : Low and Middle Income Country Profiles. Technical Report. The World Bank. URL: http://hdl.handle.net/10986/33363

World Bank, (2021). Traffic Crash Injuries and Disabilities : The Burden on Indian Society. Technical Report.TheWorldBank.URL:http://documents.worldbank.org/curated/en/761181612392067411/ Traffic-Crash-Injuries-and-Disabilities-The-Burden-on-Indian-Society.

Wright, R.E., (1995). Logistic regression, in: Grimm, L.G., Yarnold, P.R. (Eds.), Reading and Understanding Multivariate Statistics, American Psychological Association.

Zhang, G., Li, G., Cai, T., Bishai, D.M., Wu, C., Chan, Z., (2014). Factors contributing to hit-and-run crashes in China. Transportation Research Part F: Traffic Psychology and Behaviour 23, 113–124. URL: https://doi.org/10.1016/j.trf.2013.12.009, doi:10.1016/j.trf.2013.12.009.

Zhang, H., (2004). The optimality of naive bayes. Aa 1, 3. URL: https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf.

Zhang, J., Li, Z., Pu, Z., Xu, C., (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access 6, 60079–60087. doi:10.1109/ACCESS.2018.2874979.