Random forest models for motorcycle accident prediction using naturalistic driving based big data

Peer-reviewed author version

Outay, Fatma; ADNAN, Muhammad; Gazder, Uneb; Baqueri, Syed Fazal Abbas & Awan, Hammad Hussain (2023) Random forest models for motorcycle accident prediction using naturalistic driving based big data. In: International Journal of Injury Control and Safety Promotion, 30 (2) , p. 282-293.

DOI: 10.1080/17457300.2022.2164310
Handle: http://hdl.handle.net/1942/39316

RANDOM FOREST MODELS FOR MOTORCYCLE ACCIDENT PREDICTION USING NATURALISTIC DRIVING BASED BIG DATA

Fatma Outay[a], Muhammad Adnan[b] , Uneb Gazder[c*], Syed Fazal Abbas Baqueri[d], Hammad Hussain Awan[e]

[a]*College of Technological Innovation (CTI), Zayed University, Dubai*

E-mail: fatma.outay@zu.ac.ae

[b]*Transportation Research Institute (IMOB)- Hasselt University, Belgium*

E-mail: muhammad.adnan@uhasselt.be

[c]*Department of Civil Engineering, University of Bahrain, Bahrain*

[*]*Corresponding author. E-mail: ugazder@uob.edu.bh*

[d]*Department of Civil Engineering, DHA Suffah University, Karachi, Pakistan*

E-mail: fazal.abbas@dsu.edu.pk

[e] *The University of Lahore (Islamabad Campus), Islamabad, 44000, Pakistan*

E-mail: hammad.hussain@ce.uol.edu.pk

Abstract

Motorcycle accident studies usually rely upon data collected from road accidents collected through questionnaire surveys/police reports including characteristics of motorcycle riders and contextual data such as road environment. The present study utilizes big data, in the form of vehicle trajectory patterns collected through GPS, coupled with self-reported road accident information along with motorcycle rider characteristics to predict the likelihood of involvement of a motorcyclist in an accident. Random Forest-based machine learning algorithm is employed by taking inputs based on a variety of features derived from trajectory data. These features are mobility-based features, acceleration event-based features, aggressive overtaking event-based features and motorcyclists socio-economic features. Additionally, the relative importance of features is also determined which shows that aggressive overtaking event-based features have more impact on motorcycle accidents as compared to other categories of features. The developed model is useful in identifying risky motorcyclists and implementing safety measures focused towards them.

Keywords: naturalistic driving based big data; motorcycle accident prediction; random forest; machine learning; Karachi

Nomenclature

The following symbols are used in this paper:

$I_m^{T_a}$ = information about motorcyclist m in the period T_a

T_a = prediction intervals

T_t = target intervals

X^{T_a} = vector representing the behaviour of *M* motorcyclists

d_h = the duration (time span) of data that is considered to monitor motorcyclist behaviour

d_s = the duration for which probability of an accident is predicted

p_m = probability of a motorcyclist m involving an accident in the next time period

t_p = moment at which probability of an accident is predicted

$x_m^{T_a} = \langle v_1 \,,\; v_2 \,,\; v_3, \ldots, v_n \rangle$ were determined for a period T_a

$y_m^{T_t}$ = vector indicating experience of motorcyclist *m* to involve in an accident or no accident in the target period

AO_avg_acc = average acceleration during the aggressive overtaking event

AO_avg_speed = average speed during the aggressive overtaking event

AO_count = total aggressive overtaking event count

AO_max_acc = maximum acceleration during the aggressive overtaking event

F-1 score = accuracy measure for RF model

FN = false negative cases

Hacc_avg_acc = average acceleration during the harsh acceleration event

Hacc_count = total harsh acceleration event count

Hacc_dur = duration of the harsh acceleration event

Hacc_max_acc = maximum acceleration during the harsh acceleration event

Hbrk_avg_acc = average acceleration during the harsh braking event

Hbrk_count = total harsh braking event count

Hbrk_dur = duration of harsh braking the event

Hbrk_max_acc = maximum acceleration during the harsh braking event

HLacc_avg_acc = average acceleration during the harsh lateral acceleration event

HLacc_count = total event harsh lateral acceleration count

HLacc_dur = duration of the harsh lateral acceleration event

HLacc_max_acc = maximum acceleration during the harsh lateral acceleration event

N_trip = number of trips

TP = true positive cases

Introduction

A significant amount of literature has studied the causes of motorcycle accidents or factors associated with the severity of motorcycle accidents by utilising data collected from hospitals, police reports and national accident databases (Pervez et al., 2021; Waseem et al., 2019; Aidoo and Amoh-Gyimah, 2020; Wali et al., 2019; Shaheed and Gkritza, 2014). Some studies utilized other contextual data (such as traffic and weather-related data) along with road accident data (Theofilatos et al., 2018). Accident prediction modelling studies, in relation to motorcycles, have mainly used statistical approaches, however, a few of them also utilized machine learning techniques (Quddus, 2001; Harnen et al.; 2003, Ackaah and Salifu, 2011; Wahab and Jiang, 2020; Rezapour et al., 2020a; Rezapour et al., 2020b). Studies also exist that attempt to focus on real-time prediction of individual road accident. These studies used data by closely monitoring vehicles and capture events that may lead to an accident in the next few seconds, therefore, help the drivers to avoid the collision. Savino et al. (2020) reviewed such advanced technologies for two-wheelers vehicles.

The previous studies have largely shown that driver behaviour is an important, and arguably the most common, contributor to road accident (Bıçaksız and Özkan, 2016). Hence, understanding of driver characteristics and driving styles may provide valuable insights about the occurrence of accidents. However, no study was found that modelled motorcyclists involved in accidents utilising mobility data (trajectory data), which is the focus of this study. Hence, the present research adds to the present knowledge by outlining the procedure and important guidelines on utilization of big data for studying microscopic characteristics of motorcyclists, including their risk-taking behaviour. Moreover, it also provides the features, related to driving styles, which may contribute to the involvement of motorcycle drivers in accidents.

This article is structured as follows: Section 2 discusses the practical application of this research. Section 2 gives an overview of the studies conducted so far in this domain. Section 3 explains collected data and machine learning models used to predict the risk of involving in a road accident in the near future. Section 4 provides details of experimental settings, presents the obtained results and also provides a detailed discussion on model outcomes. Section 5 is conclusion of this research study.

Practical Application

This work can prove to be important for motorcycle insurance companies, as the motorcyclists' risk of having accidents in the near future is modelled. Motorcyclists that may have a higher risk are likely to cause more burden compared to the ones that have a lower risk. The analysis done in this study is expected to allow insurance companies to develop individual-specific pricing policies, with a trade-off between profit and competitiveness. It is also expected to provide the basis, to the authorities, for identifying risky motorcycle drivers and employing measures to curtail their risk-taking behaviour. Because of the several advantages of motorcycles as a transport mode, there are several new businesses who have set their premise on engaging private motorcyclists or by managing a fleet of motorcycles to provide specific services (delivery, ride-hailing etc.) to their customers. Some notable businesses in Pakistan are MyRider (https://www.myrider.pk/) and Bykea (https://www.bykea.com/). Therefore, this research explores the avenue of using trajectory data for risk assessment. Such data is already been collected by the fleet management companies for cars and motorcycles.

Literature review

The majority of the studies used statistical models in order to relate motorcycle accidents with a variety of explanatory variables. Some of the statistical models used in these studies are as follows: log-linear model, logistic regression, mixed ordered logit model, ordered probit, random parameter logit and structural equation modelling (SEM) (Radin et al., 2000; Lam et al., 2019; Chang et al., 2016; Cunto et al., 2017; Seva et al., 2013; Chung et al., 2014; Waseem et al., 2019; Pervez et al., 2021; Bathan et al., 2018). A few studies used machine learning models such as multi-layer perceptron, rule induction and classification and regression tree (Wahab and Jiang, 2020; Adnan and Gazder, 2019; Rezapour et al. 2020a). Rezapour et al. (2020b) also employed a deep neural network to analyse motorcycle accident severity. These recent studies advocated the use of machine learning models because of their better predictive performance compared to parametric models. Additionally, in the era of big data, a large variety of data is available and machine learning methods can help in analysing complex non-linear relationships.

In the context of Karachi, a number of studies have found the high involvement of motorcycle riders in road accidents. Hassan et al. (1997) reported that victims of these accidents were young (mean age of 31 years) and constituted almost 50% of total resulting injuries and fatalities. Lateef (2010) attributed their dominance in road accidents to their significantly high presence in the traffic stream and changes in infrastructure leading to high-speed signal free corridors. Mirza et al. (2013) found that motorcycle rider fatalities in road accidents were mainly due to head and chest injuries. In addition to that, Martins et al. (2021) also reported multiple bone injuries resulting from motorcycle accidents as well.

The use of vehicle trajectory data was initially advocated for studying the microscopic parameters of vehicles, such as car-following, lane-changing, etc. This type of data is more commonly extracted from high-resolution images (Kovvali et al., 2007). Punzo et al. (2011) has also advocated its use for simulation programs to understand and calibrate the parameters of traffic flow theory. Another advancement in this field is the use of GPS navigation data, coupled with road network topology. This approach compensates for the lower sampling rate of trajectory which is collected through GPS (Liu et al., 2012). Drone technology has also been utilized for collection of vehicle trajectory data recently (Zhong et al., 2020). Heterogeneity in traffic always poses a challenge for traffic analysts, especially in case of observing microscopic traffic phenomena such as car following. High resolution vehicle trajectory data has been used efficiently to update the traditional models for heterogeneous traffic (Taylor et al., 2015). Other applications of this vehicle trajectory data include; optimization of traffic signals (Ma et al., 2020). Khekare et al. (2022) used the trajectory data, collected from a roundabout, to calculate the capacity of a roundabout. They employed computer vision and machine learning technique in their analysis. They emphasized on the use of such data for roundabout due to the nature of traffic flow which is largely dependent upon driver decisions, making it more vulnerable to accidents.

Considering effectiveness of vehicle trajectory in observing and predicting vehicular behaviour, researchers have also used it for prediction of accidents (Oh and Kim, 2010). Park et al. (2018) proposed the use of this data for analysing lane change risk index, consequently, modelling the accident potential. Wang et al. (2019) has also utilized the video-based trajectory data for accident prediction.

Almost in all the studies cited above, the investigations are based on the availability of accident data which is coupled with some contextual information such as motorcyclist’s characteristics (age, motorcycle engine capacity, education, driving experience etc.), road environment at accident location (i.e. type of road features such as intersection, roundabout, number of lanes, road type, speed limits, traffic flow, weather condition etc.) and other variables such as temporal aspects. Among the motorcyclist’s characteristics, age, and driving experience have been found in earlier studies as major contributing factors in involvement in accidents (Pervez et al., 2021). Speeding and helmet wearing behaviour have been found as factors that significantly correlate with severity of motorcycle accidents (Waseem et al., 2019). Motorcyclist's aggressive behaviour has been advocated in the literature as a strong predictor of involvement in accidents (Tasca et al., 2000), however, collecting such data is difficult. Bathan et al. (2018) collected driver angriness, self-assertiveness and rule violation information using the questionnaire survey and then employed these variables along with others in an SEM framework. Chung et al. (2014) used a video image detector for the available CCTV footages just before the time of the accident to determine the speeding behaviour of the motorcyclists. Furthermore, obtaining speed just before the accident, can provide some idea of aggressiveness from the motorcyclists, but to appropriately ascertain about aggressive and risky driving behaviour, such data need to be collected for a longer period. There is no study found, where such data is used which is not self-reported but collected through observation. Processing trajectory data collected over a longer period of time can be very useful as it provides detailed insight into driving behaviour.

To this end, this study advances the current literature by incorporating trajectory data into the analysis of accident prediction and uses a robust algorithm available within the machine learning models to correlate trajectory features with the involvement of accidents. Additionally, features that correspond to aggressive behaviour of the motorcyclist were also extracted from the trajectory data and incorporated in the developed model.

Data and methods:

Data collection, preparation and its characteristics

GPS trajectory data used in this study was collected from motorcyclists in Karachi (Pakistan). A device was used that can easily be mounted on the motorcycle. The device includes a SIM card and connects via a GSM network to transmit the data to the server. Under a Motorcyclist Safety Programme, 2000 motorcyclists were recruited for this study that belonged to people from different age groups and also involved in different professions. The recruitment was done on voluntary basis without any intervention from the traffic police or any other government organization. Therefore, the distribution of data, in terms of motorcyclists' characteristics and accident occurrence, may not be true of the representative of the population. However, the trends shown in sample between accident occurrence and other parameters can still be considered valid.

To reduce the cost burden of the GPS devices (total devices used in the study were approximately 550), the recruitment was done in four batches in a manner that for at least five (05) consecutive months data can be collected from each motorcyclist. Only those motorcyclists were recruited, where there is a high likelihood that the person recruited is the only driver that will ride on the motorcycle for at least the next five months. This exercise was planned to be completed in 18 months, starting in January 2019 and it was scheduled to be completed in June 2020. However, due to the COVID-19 pandemic, a country-wide lockdown was initiated in the month of April 2020. So the data from the motorcyclists collected until March 2020 was utilized. Along with the GPS based trajectory data, information about individual socio-economic status was also collected. Further to that information, all recruited participants during the data recording period were advised to provide details of the accident (by filling the short form and send us back via e-mail) in which they may have been involved while driving the motorcycle.

All the participants were given confidence in the secrecy and protection of the data and they also had the flexibility to drop out from the study based on approved ethical requirements. Participants were also given an incentive in the form of a voucher, which can be availed at fuel stations to have a 20% reduced cost of fuel price. This incentive is based on the usage of motorcycle, which can be monitored through positioning data to avoid fraudulent situations.

Naturalistic driving based big data

The data was collected from the GPS tracker device at an average rate of 4.9 points per minute and it contained rich information such as positioning in terms of latitude and longitude, time, vehicle speed, and driving events (such as harsh acceleration, harsh brake, harsh lateral acceleration). The device constantly generates data and its frequency of data generation increases while the vehicle is in motion, and also when an event occurs. The dataset used in this study consists of data of 1,612 motorcyclists, with a total size of around 70GB. The cleaning of the dataset was performed by following Adnan et al. (2020), such as removal of outliers (positioning points) associated with a very high instantaneous vehicle speed (i.e. speed more than 100km/hr). Speeds of motorcycles range between 15-80 km/hr, it should be noted that types of motorcycles prevailing in Pakistan are mostly equivalent to *Honda CD 70 model* (manufactured locally and also imported from China) that have maximum speed up to 120km/hr. Duplicate position points that have the same time stamps were also removed. In addition to this, the distance between the two consecutive points is also measured and found in a range of 0 – 250m, with a typical sampling rate of around 0.2 minutes which corresponds to the speed of 0 and 75 km/hr respectively. Trips from the trajectory data were extracted by using a temporal gap of 15minutes between the stop points. At the same time, our trip extraction algorithm ensures that when position remained within a small area (instead of only at a stopping point) during a time interval of 15 min, that small area is treated as a stopping point to account for any measurement errors. Total trips extracted from the dataset is around 0.247 million, this gives an average number of trips per person per day as 2.7 from 1,612 motorcyclists. This figure is also reasonable as the average trips per person reported in the literature for Karachiites are roughly around 2.1 trips per day (Hasan & Raza, 2015). Motorcyclists usually have a high tendency to perform more trips because of the higher flexibility and accessibility of this transport mode.

For extracted trips, the data was prepared for analysis considering various types of features such as trip-based or mobility-based features and event-based features. Mobility-based features are characterizing a user based upon classic indicators of the trips and aspects that describe them such as length (km), duration (seconds) and speed (km/hr). Event-based features include detected events from the device and also from the processing of trip trajectories. Acceleration based events that are detected by the device such as harsh acceleration, harsh braking, and harsh lateral acceleration are included in the analysis. For each such events, the available average acceleration magnitude (km/sec^2), its duration (sec), the maximum value of acceleration and their counts are determined. Additionally, trip trajectories are further processed to capture an event which is a special characteristic of motorcycle driving behaviour as they usually do not observe lane discipline. This is termed as zig-zag behaviour or aggressive overtaking behaviour of the driver to manoeuvre their vehicle in a moderate to high congested environment (Minh et al., 2012; Halim et al., 2020). An imprecise method was used to detect such events, within which trip trajectories are segmented into 300-400 m in length (distance covered while traversing the trajectory) and then for each segment, the standard deviation in the lateral positions (SDLP) are observed provided that at least three position points are available. If the value of SDLP exceeds a threshold value (also taking account of measurement error) of 25 cm (Verster and Roth, 2011), then the event has occurred otherwise not. The illustration shown in Figure 1 explain this further. In one part of the trajectory segment where distance covered is 383m, the SDLP exceeds the threshold value, and therefore this type of event is occurred.

<Insert Figure 1 here>

For this event; count, average speed, average acceleration and maximum acceleration are further determined. These indicators are aggregated through four operators; counts, sums, means and standard deviations (wherever relevant). Additionally, these aggregated operators are computed over four time periods of the day: morning (6am – 12noon), afternoon (12noon – 5pm), evening (6pm – 10pm) and night (10pm -6am). Table 1 describes the summary of these features that are extracted from trajectory data.

<Insert Table 1 here>

Motorcyclist information

During the recruitment phase, individuals were asked to provide socio-economic information to keep up with a variation that exists in the motorcyclist driving population in Karachi, Pakistan. Individuals have provided information on their *age*, *marital status*, *driving experience*, *education*, *occupation*, *income*, *number of vehicles (car + motorcycle) availability in the household*, *presence of children (under 15 years) in the household*. It should be noted that due to cultural norms, females are not driving motorcycles in Pakistan, therefore, the *gender* variable is not included in the analysis. Table 2 provides more insight on socio-economic data and the % frequency of sample data. In the absence of population statistics for motorcyclists, the categories for income and age were set as per the findings of previous studies, including (Adnan and Gazder, 2019; Ali, et al., 2021; Marvi et al., 2022). Moreover, sample frequency distribution of age, household income, marital status and income level was compared with other similar studies for Karachi motorcyclist and a reasonable match was found.

<Insert Table 2 here>

Motorcyclist accident data

This data was collected in a self-reported manner. Recruited motorcyclists were given a digital form that asks the individual to provide information on any type of road accident that occurred while driving a motorcycle. Motorcyclists or their close relatives can e-mail us this information. Motorcyclists were also given an option to report this accident on a particular phone number in case they do not have access to digital services (internet/e-mail etc.). It was not necessary to report immediately after the accident, however, the information that came as quickly as possible was checked through other sources to register a particular accident with more credibility. Table 3 provides insight into collected information and also reports the collected aggregate statistics. In total, 687 accidents occurred with the recruited motorcyclists, the majority of which are of minor injury or property damage type. Only 1 fatal and 15 major injury accidents were reported. It should be noted that out of these 687 accidents, 98 were reported to police (shows that the police record is highly underestimating the number of accidents) and 201 were those in which victims were transferred or visited hospitals to recover from injuries. For this study, authors were given access to the location and time of the accident data only (687 location points with their timestamp) with the id of individuals to match them with their personal characteristics and trajectory data.

<Insert Table 3 here>

Problem formulation and modelling method

The problem of accident prediction was defined as inferring probability of having an accident in the next time period (next month) for each motorcyclist (m). This is similar to the problem formulated by Guidotti and Nanni (2020) for car accidents using their trajectory data, as shown in Equation (1). In relation to that, two time intervals were defined, namely; T_a and T_t named as prediction and target intervals. These intervals are given as follows:

$$T_a = [t_p - d_h \, , t_p] \text{ and } T_t = (t_p \, , t_p + d_s] \tag{1}$$

where, t_p is the moment at which probability of an accident is predicted, d_h is the duration (time span) of data, that is considered to monitor motorcyclist behaviour (such as past 3 months) and d_s is the duration for which probability of an accident is predicted (i.e. a month). The probability p_m of a motorcyclist *m* involving an accident in the next time period is given by equation (2).

$$p_m = P\ (m\ has\ a\ crash\ in\ \mathrm{T}_t \mid I_m^{Ta}\ ) \tag{2}$$

where, I_m^{Ta} is an information about motorcyclist m in the period T_a, obtained through features of trajectory data, accident information (accident data) and socio-economic data. The probability is modelled using standard machine learning algorithms for a binary outcome (accident/no accident). For each motorcyclist *m,* a vector of their features $x_m^{Ta} = \langle v_1\,,\ v_2\,,\ v_3, \dots, v_n \rangle$ were determined for a period T_a . A vector X^{Ta} was prepared representing the behaviour of *M* motorcyclists. Another vector $y_m^{T_t}$ was defined indicating experience of motorcyclist *m* to involve in an accident or no accident as (1 or 0) in the target period. Corresponding to $y_m^{T_t}$, a vector Y^{T_t} was defined for M motorcyclists.

Machine learning classifiers can be trained by providing inputs X^{T_a} and Y^{T_t}, that provides output in the form of probability of motorcyclist being involved in an accident p_m. In this study, K-NN classifier, support vector machine, decision trees and random forest (RF) methods were tried as machine learning classifiers. Based on some earlier attempts, RF method was found to be the most accurate. It was also found to be consistent in terms of providing similar performance on training and validation datasets. These results are also validated by previous studies which have found RF models to give better performance in comparison to other models (Xing et al., 2019; Sun et al., 2021). Additionally, it has been also reported in the literature that among all ensemble methods, RF has shown good capability in solving prediction and classification problem (Zaklouta and Stanciulescu, 2012; Zhang and Haghani, 2015; Cheng et al., 2019). This is because RF combines multiple simple decision trees to optimize predictive performance instead of a single best tree model. Due to these reasons, the results reported in the next section are described for the random forest method only.

<Insert Figure 2 here>

The actual working of the algorithm is shown in Figure 2. The bootstrap sampling procedure split the training data set randomly with replacement into K samples (a parameter for calibration), based on which K base decision trees were developed during the model estimation phase. Additionally, to avoid correlation between the base decision trees, random feature selection was also employed at this stage which ensures that not all explanatory variables are used but rather a random subset of them will be used at each splitting node. The number of splitting variables (N_s) is also a parameter for calibration. Another important parameter for calibration is the maximum tree depth (D_t). In short, two levels of randomness are employed in the algorithm i.e. random splitting of the training set with the same sample size and a different set of explanatory variables to split each node. In every single tree, splitting continues until maximum depth is reached for the tree. Once base tree models are estimated, the majority voting strategy is used for results computation. The final outcome is the one that is predicted most across the ensemble.

Results and discussion

Experimental settings

The preparation of the dataset follows what has been mentioned in section 3.1. The dataset based on motorcyclist is divided into two parts such as training and testing datasets with a ratio of 70:30 respectively. To apply a machine learning classifier to the training dataset, it was first needed to resolve the issue of classes' imbalance, since there is an inherent assumption of almost all machine learning classifiers that the number of examples for each class should be equal. This is not the case with the problem in this study as there is a significant imbalance between the accident and no accident in the training dataset. As a solution to this issue, authors applied the technique originally proposed in Chawla et al. (2002) and known as SMOTE oversampling approach. Synthetic examples were introduced along the line joining the k minority class nearest neighbours. k=5 was used, as recommended in Chawla et al. (2002). This improves the class balance and also ensured the presence of the minority class in the decision regions where it appears. It should be noted that only the training dataset was re-balanced and not the testing which makes the evaluation harder but more appropriate. Furthermore, as various features of the motorcyclist were prepared as a part of vector X^{T_a}, to ensure that features are independent of each other, correlation analysis (Pearson correlation co-efficient) was performed among the pair of features and identified highly correlated features (i.e. correlation co-efficient ≥ 0.7). During the model training, a single feature, out of those which were found highly correlated, was used to avoid the effect of double-counting a particular feature. In total, around 89 features were used for training the model. Additionally, experiments were done to obtain the optimum values of K, N_s and D_t based on the process described in Cheng et al (2019). The calibrated model utilizes K= 350, N_s = 6 and D_t = 1600 and resulted in a model that is more robust and achieves better accuracy with optimum use of computational resources.

Evaluation measures and model interpretation

Given the objective of the study that required accurate prediction of accidents (harmful events), and also the imbalanced nature of the dataset, it is important to analyse the measures that signify the extent of false negatives rather than false positives. Recall in relation to positive class (minority class i.e accident occurrence) would give insights on the model performance to identify as many risky motorcyclists as possible. Precision in relation to negative class (majority class, i.e. no accident) would give insights on the model performance to raise no alarm only if modeller is confident that motorcyclist is not risky. Furthermore, f1 score in relation to positive class is also useful as it takes both recall and precision into consideration. All of these measures range from 0 to 1, with 1 being associated with the most accurate model. They were calculated using the equations (3) to (5).

$$Precision = TP/(TP+FP) \tag{3}$$

$$Recall= TP/(TP+FN) \tag{4}$$

$$F1\ Score = (2*Precision*Recall)/(Precision+Recall) \tag{5}$$

Where, *TP* refers to true positive cases wherein a accident occurred and the model predicted it to occur, and *FN* refers to a false negative cases wherein the accident occurred but the model predicted not to occur (Goutte and Gaussier, 2005).

Table 4 present statistics on these measures for the test dataset. The value of these evaluation measures represents that the model is able to classify the two imbalance classes considerably well.

<Insert Table 4 here>

Another important result is available in the form of features' importance as shown in Table 5 (Only the top 25 features are mentioned) and Figure 3. The random forest method utilizes a *MeanDecreaseGini* as a measure of variable importance based on the Gini impurity index used for the calculation of splits during training (Atkinson, 1970). Gini importance indicates how often a particular feature was selected for a split, and how large its overall discriminative value was for the classification problem under study. Equivalently, ranking of features based on their importance reflect that in a given model the higher rank features are most important in explaining the target variable. To obtain stable importance of features, the model was ran around 80 times with the training dataset.

<Insert Figure 3 here>

<Insert Table 5 here>

Discussion

Based on the model performance statistics shown in table 4, the model seems to be performing reasonably well, given the imbalanced nature of the data. Authors also performed 5-fold cross-validation and the results of that validation were also similar to what is shown in Table 4. Table 5 provides useful insights on which features explain the occurrence of an accident in the near future. These top 25 features share the total relative importance of around 66%. Table 5 clearly illustrates that all four types of features incorporated in the model explain the accident occurrence near future. Features related to the aggressive overtaking behaviour of motorcyclists seem more important as on average the relative importance of these features was found around 3.9% (among top 25 features). Moreover, these features tend to increase the likelihood of the motorcyclist being involved in a crash. Within these features, total count of AO events is at the top. Additionally, the existence of features (among the top 25) such as average acceleration during these events especially in morning and evening times (which are usually rush hours and during these time queue phenomena is also present) also indicates risk-seeking driving behaviour and therefore, may enhance the chances of involvement in an accident in the near future. Among the motorcyclist's socio-economic characteristics, age and driving experience are found to be better predictors, though education, household size and household income are also present in the top 25. Household size is an important feature as it reflects the cultural notion that often motorcycle trips are joint trips with other household members (children, spouse etc.) as pillion riders. Extra load on the motorcycle in the form of family and friends may exhibit cautious behaviour from motorcyclist that may reduce chances in the involvement of accidents, as shown in Figure 4. On the other hand, age seems to be inversely proportional to frequency of accidents, as shown in Figure 5. Within the category of features related to acceleration-based events, total count of harsh braking and harsh acceleration, and average acceleration in the morning and evening times have higher importance. These features also relate to erratic driving behaviour and therefore may enhance the chances of occurrence of an accident in the near future.

These results were compared from the findings mentioned in the literature. Though as mentioned previously, there was no study found that used trajectory data to predict motorcycle accidents, however, similar study exists for car accidents in the Tuscany (Italy), Rome (Italy) and London (UK) regions, where they utilised trajectory data coupled with accident data (Guidotti and Nanni, 2020). They reported in their study that the event- and mobility-based features improved the model performance. Accident investigation literature often used statistical models to assess the role of different factors that may have caused the accidents. Bathan et al. (2018) showed that self-assertiveness, anger, speeding and rule-violation exhibit a direct relationship with the involvement of accidents through structural equation modelling. They collected data on personality trait with the help of a questionnaire survey. With the help of big trajectory data, to some extent personality trait, such as self-assertiveness can be reflected through AO based events in the present study, in which motorcycle riders follow a zig-zag manoeuvre with high acceleration. By doing this, motorcyclists eventually attempt to assert their skills in carrying out such manoeuvres in rush hours. RF model calibrated in this study, gives higher importance to AO-based features and therefore results of this study are aligned with the findings of Bathan et al. (2018). Moreover, this study provides a way to quantify such self-assertiveness directly through the observed data. Along with aggressive overtaking, acceleration-based events such as harsh acceleration, harsh braking and harsh lateral acceleration have been found important features to explain the prediction of accidents. These features reflect the aggression in motorcyclists. These are also found significant in some previous studies such as Halim et al. (2020), Waseeem et al. (2019), Seva et al. (2013) and Tasca et al. (2000). Pervez et al. (2021), in their study, found out that among motorcyclists' socio-economic characteristics, age and pillion rider presence significantly affect their involvement in fatal accidents. The RF model calibrated in this study also shows the importance of these features in the involvement of accidents in near future. The rank of age was found as 3 (in terms of importance) and household size (that may reflect the tendency of a higher number of trips with other family members) stands on rank 15.

<Insert Figure 4 here>

<Insert Figure 5 here>

Additionally, in relation to temporal characteristics, Pervez et al. (2021) mentioned that morning and night hours are more critical in terms of the occurrence of fatal accidents. The results mentioned in table 5 support these findings as time-based mobility features such as *Average_morning_trips*, *Average_morning_speed*, *Total_night_distance* are present among the top 25 features. Additionally, event-based features such as *Average_morning_Hbr_avg_acc* and *Average_evening_Hbr_avg_acc* are also good predictors of involvement of accidents in near future. The likelihood of crashes seem to increase proportionally with *Average_morning_trips* and *Total_night_distance*. In order to reach to their work location during morning rush hour, motorcyclists tend to exhibit risky behaviour and in the night time, the chance of accidents are higher due to poor lighting and inappropriate road condition (such as potholes, uneven surface and poor drainage etc.). Other studies such as Rifat et al. (2012), Seva et al. (2013), Manan et al. (2018) and Vajari et al. (2020) reported similar findings on temporal characteristics.

One of the key distinguishing factors of this study is the use of big trajectory data and its processing to define mobility, acceleration-based events and aggressive overtaking features for utilizing them in the prediction of a motorcycle accident in the near future. Because of the use of big trajectory data, employing a machine learning algorithm in the form of random forest is also natural as the capability of RF in handling different types of variables and modelling complex nonlinear relationships makes it a promising method. Additionally, the problem of accident prediction formulated in this study also has significant importance for motorcycle insurance companies and smart startups that manage the fleet of motorcycles. The model developed in this study can be used to devise a range of insurance policies (pricing and possible benefits) that have their premise on dynamic changes in motorcycle driving behaviour to attract a larger customer base. Fleet managers based on the developed model results can identify risk-seeking drivers and developed programs/intervention to nurture their driving behaviour in the right direction. In the short run, the future work will involve further processing of the trajectory data to obtain more mobility features that are relevant to space and time. Individual mobility network (IMN) presented in (Guidotti and Nanni, 2020) and features derived from these networks can be the next possible candidate. In the long run, an automation pipeline, where such trajectory data is collected in real-time, model improvement and monitoring risky driving behaviour can provide an alert mechanism to timely inform such risk-seeking motorcyclists of the danger they may cause to themselves and others.

Conclusion

This research aimed at utilizing big related to vehicle trajectory to study the risk-taking behaviour of motorcycle riders in Karachi. The data was utilized in RF models for prediction of accidents.

Based on the analysis and discussion presented in Sections 4, the RF model developed from the collected big trajectory data from motorcyclists revealed that driving behaviour is a significantly important notion that explains accident prediction. Trajectory data seems very fruitful in deriving relevant features that reflect aggressive driving behaviour and driver exposure (features that represent per day trips, travel distance and travel duration). The study also showed that along with the trajectory data, socio-economic characteristics are also important and these should not be ignored in the accident prediction problem. The model shows that factors related to aggressive driving behaviour, including AO and average acceleration increase the likelihood of accidents. Larger family size and age are shown to have a negative impact on the likelihood of accidents. Driver exposure seems to increase the likelihood of crashes, which is shown by the distance travelled, number of trips and harsh braking and acceleration events in the morning and evening peak hours.

RF algorithm, as shown in various previous studies, stands out among the available machine-learning algorithms to tackle the imbalance accident prediction problem. The problem formulation and model results presented in this study create new avenues for insurance companies to introduce insurance plans and their pricing in relation to the driving behaviour of motorcyclists. At the same time, the developed model provides opportunities for motorcycle fleet managers to identify risky motorcyclist and therefore can further nurture their skills and behaviour through relevant intervention programs.

In light of the findings of this study, it is recommended that focused policies should be implemented to penalize drivers with risky behaviour. This could involve introducing traffic-violation points system wherein drivers with repetitive violations receive points which are part of their driving record. It is shown that this behaviour more often exhibited in the peak hours, hence, it is expected that better traffic management and enforcement strategies may reduce such behaviour. These strategies may include restricting access to main highways and exclusive lanes.

There were two limitations of this study, which could be incorporated in future research. Firstly, the absence of official statistics related to motorcycle riders in Karachi. Secondly, the inability of the motorcyclists to record precise accident location as they were part of the accidents.

It is also recommended for future studies to incorporate a larger population of motorcyclists for observation of trajectory and accidents, which may require government intervention. Secondly, similar data can be collected and utilized for developing models to predict accident severity.

Data availability statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. Vehicle trajectory data, driver characteristics data, random forest model code.

Acknowledgements

This research was supported by Zayed University Research Cluster grant #R17075. Authors also acknowledge the support provided by Mr. Hakeem Ahmed (Manager, Motorcycle Safety Program, Karachi, Pakistan), to provide access to trajectory data of motorcyclists along with their relevant information and accident statistics.

Disclosure statement

The authors report there are no competing interests to declare.

References

Ackaah, W., & Salifu, M. (2011). Accident prediction model for two-lane rural highways in the Ashanti region of Ghana. *IATSS Research* 35(1):34-40.

Adnan, M., & Gazder, U. (2019). Investigation of helmet use behavior of motorcyclists and effectiveness of enforcement campaign using CART approach. IATSS Research 43(3):195-203. DOI: 10.1016/j.iatssr.2019.02.001.

Adnan, M., Gazder, U., Yasar, A., Bellemans, T, & Kureshi, I. (2020). Estimation of travel time distributions for urban roads using GPS trajectories of vehicles: a case of Athens, Greece. Personal and Ubiquitous Computing. https://doi.org/10.1007/s00779-020-01369-4

Aidoo, E. N., & Amoh-Gyimah, R. (2020). Modelling the risk factors for injury severity in motorcycle users in Ghana. Journal of Public Health 28:199–209. https://doi.org/10.1007/s10389-019-01047-7

Ali, A., Malik, M. A., Khan, U. R., Khudadad, U., Raheem, A., & Hyder, A. A. (2021). Helmet wearing saves the cost of motorcycle head injuries: a case study from Karachi, Pakistan. *ClinicoEconomics and outcomes research: CEOR*, *13*, 573.

Bathan, A., de Ocampo, J., Ong, J., Gutierrez, A., Seva, R., & Mariano, R. (2018). A predictive model of motorcycle accident involvement using structural equation modeling considering driver personality and riding behavior in Metro Manila. *Proceedings of the International Conference on Industrial Engineering and Operations Management,* 2018-March:1783-1804.

Bıçaksız, P., & Özkan, T. (2016). Impulsivity and driver behaviors, offences and accident involvement: A systematic review. *Transportation research part F: traffic psychology and behaviour*, *38*, 194-223.

Chang F, Li, M. Xu, P., Zhou, H., Haque, M. M., & Huang, H. (2016). Injury severity of motorcycle riders involved in traffic accidentes in Hunan, China: a mixed ordered logit approach. *Int J Environ Res Public Health* 13:E714. pmid:27428987

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321-357.

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society* 14:1-10.

Chung Y., Song, T. J., & Yoon, B. J., (2014). Injury severity in delivery-motorcycle to vehicle crashes in the Seoul metropolitan area. *Accid Anal Prev.* 62:79–86. pmid:24161584

Cunto F. J. C., & Ferreira, S. (2017). An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. *J Transp Saf Secur* 9:33–46.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*. 2005. Springer.

Guidotti, R., & Nanni, M. (2020). Accident prediction and risk assessment with individual mobility networks. *21st IEEE International Conference on Mobile Data Management (MDM)*, Versailles, France, 2020: 89-98, doi: 10.1109/MDM48529.2020.00030.

Halim, H., Mustari, I., & Saing, Z. (2020). Overtake behavior model of motorcycle on accident risk. *International Journal of Advanced Science and Technology, Science and Engineering Research Support Societ* 29 (5):12556-12567. ff10.5281/zenodo.3908682ff. ffhal-03090270

Harnen, S., Wong, S. V., Umar, R. R., & Hashim, W. W. (2003). Motorcycle accident prediction model for non-signalized intersections. *IATSS Research* 27(2):58-65.

Hasan, A., & Raza, M. (2015). Responding to the transport crisis in Karachi. *IIED and Urban Resource Center. See: http://pubs. iied. org/10733IIED. html.*

Khekare, P., Bonthu, S., Hunt, V., Helmicki, A., & Lee, K. (2022). A case study on multilane roundabout capacity evaluation using computer vision and deep learning. *Journal of Computing in Civil Engineering* 36(3):05022001.

Kovvali, V. G., Alexiadis, V., & Zhang PE, L. (2007). *Video-based vehicle trajectory data collection* (No. 07-0528).

Lam C, Pai, C-W., Chuang, C-C., Yen, Y-C., Wu, C-C., Yu, S-H., et al. (2019). Rider factors associated with severe injury after a light motorcycle accident: A multicentre study in an emerging economy setting. *PLoS ONE* 14(6): e0219132. https://doi.org/10.1371/journal.pone.0219132

Lateef, M. U. (2011). Spatial patterns monitoring of road traffic injuries in Karachi metropolis. *International Journal of Injury Control and Safety Promotion* 18(2):97-105.

Liu, S., Liu, C., Luo, Q., Ni, L., M., & Krishnan, R. (2012, July). Calibrating large scale vehicle trajectory data. In *2012 IEEE 13th International Conference on Mobile Data Management:*222-231. IEEE.

Ma, W., L. Wan, C. Yu, L. Zou, and J. Zheng. 2020. Multi-objective optimization of traffic signals based on vehicle trajectory data at isolated intersections. *Transportation Research Part C: Emerging Technologies* 120:102821.

Manan, A. M. M., Várhelyi, A., Çelik, A. K., & Hashim, H. H. (2018). Road characteristics and environment factors associated with motorcycle fatal accidentes in Malaysia. *IATSS Research*, 42(4):pp. 207–220.

Martins, R. S., Saqib, S. U., Gillani, M., Sania, S. R. T., Junaid, M. U., & Zafar, H. (2021). Patterns of traumatic injuries and outcomes to motorcyclists in a developing country: A cross-sectional study. *Traffic Injury Prevention* 22(2):162-166.

Marvi, H., Soomro, M., & Memon, I. A. (2022). Influence of Socio-economic Factors on Mode Choice of Employees in Karachi City. *Global Economics Review*, *7*(2), 13.

Minh, C.C., Sano K., & Matsumoto. S. (2012). Maneuvers of motorcycles in queues at signalized intersections. *J. Adv. Transp.* 46:39-53. https://doi.org/10.1002/atr.144

Mirza, F. H., Hassan, Q., & Jajja, N. (2013). An autopsy-based study of death due to road traffic accidents in metropolis of Karachi. *JPMA* 63(2):156-60.

Oh, C., & Kim, T. (2010). Estimation of rear-end accident potential using vehicle trajectory data. *Accident Analysis & Prevention 42*(6):1888-1893.

Park, H., Oh, C., Moon, J., & Kim, J. (2018). Development of a lane change risk index using vehicle trajectory data. *Accident Analysis & Prevention* 110:1-8.

Pervez, A., Lee, J., & Huang, H. (2021). Identifying factors contributing to the motorcycle accident severity in Pakistan. *Journal of Advanced Transportation 2021*. DOI: 10.1155/2021/6636130

Punzo, V., Borzacchiello, M. T. & Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C: Emerging Technologies* 19(6):1243-1262.

Quddus, M. A., Chin, H. C. & Wang, J. (2001). Motorcycle accident prediction model for signalised intersections. *WIT Transactions on the Built Environment* 52.

Radin U. R. S., Mackay, M., & Hills, B. (2000). Multivariate analysis of motorcycle accidents and the effects of exclusive motorcycle lanes in Malaysia. *Journal of Accident Prevention and Injury Control* 2(1):11-17, DOI: 10.1080/10286580008902549

Rezapour, M., Molan, A. M., & Ksaibati, K. (2020a). 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(2):89-99.

Rezapour, M., Nazneen, S., & Ksaibati, K. (2020b). Application of deep learning techniques in predicting motorcycle accident severity. *Engineering Reports*, 2(7):e12175.

Rifaat, S. M., Tay, R., & De Barros, A. (2012). Severity of motorcycle crashes in Calgary. *Accident Analysis & Prevention* 49:44–49

Savino, G., Lot, R., Massaro, M., Rizzi, M., Symeonidis, I., Will, S., & Brown, J. (2020). Active safety systems for powered two-wheelers: A systematic review. *Traffic Injury Prevention* 21(1):78-86.

Seva, R. R., Flores, G. M. T., Gotohio, M. P. T., Paras, N. G. C. (2013). Logit model of motorcycle accidents in the Philippines considering personal and environmental factors. *International Journal for Traffic & Transport Engineering* 3(2).

Shaheed, M. S. B., Gkritza, K., Zhang, W., & Hans, Z. (2013). A mixed logit analysis of two-vehicle accident severities involving a motorcycle. *Accident Analysis & Prevention*, 61:119-128.

Sun, D., Xu, J., Wen, H., & Wang, D. (2021). Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. *Engineering Geology*, *281*, 105972.

Tasca, L. (2000). *A review of the literature on aggressive driving research*. Ontario, Canada: Ontario Advisory Group on Safe Driving Secretariat, Road User Safety Branch, Ontario Ministry of Transportation. Accessed online at 30th March 2021, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.957&rep=rep1&type=pdf

Taylor, J., Zhou, X., Rouphail, N. M., & Porter, R. J. (2015). Method for investigating intradriver heterogeneity using vehicle trajectory data: A dynamic time warping approach. *Transportation Research Part B: Methodological* 73:59-80.

Theofilatos, A., & Ziakopoulos, A. (2018). Examining injury severity of moped and motorcycle occupants with real-time traffic and weather data. *Journal of Transportation Engineering, Part A: Systems* 144(11):04018066.

Vajari, M. A., Aghabayk, K., Sadeghian, M., & Shiwakoti, N. (2020). A multinomial logit model of motorcycle accident severity at Australian intersections. Journal of Safety Research 73:17–24.

Verster, J. C., & Roth, T. (2011). Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (SDLP). *International Journal of General Medicine* 4:359–371. https://doi.org/10.2147/IJGM.S19639

Wahab, L., & Jiang, H. (2020). Severity prediction of motorcycle crashes with machine learning methods. *International Journal of Crashworthiness* 25(5):485-492. DOI: 10.1080/13588265.2019.1616885

Wali, B., Khattak, A. J., & Ahmad, N. (2019). Examining correlations between motorcyclist's conspicuity, apparel related factors and injury severity score: Evidence from new motorcycle accident causation study. *Accident Analysis & Prevention* 131:45-62.

Wang, C., Xu, C., & Dai, Y. (2019). A accident prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. *Accident Analysis & Prevention* 123:365-373.

Waseem, M., Ahmed, A., & Saeed, T. U. (2019). Factors affecting motorcyclists' injury severities: an empirical assessment using random parameters logit model with heterogeneity in means and variances. *Accident Analysis & Prevention* 123:12–19.

Xing, J., Wang, H., Luo, K., Wang, S., Bai, Y., & Fan, J. (2019). Predictive single-step kinetic model of biomass devolatilization for CFD applications: A comparison study of empirical correlations (EC), artificial neural networks (ANN) and random forest (RF). *Renewable Energy*, *136*, 104-114.

Zaklouta, F., & Stanciulescu, B. (2012). Real-time traffic-sign recognition using tree classifiers. in *IEEE Transactions on Intelligent Transportation Systems* 13(4):1507-1514, Dec. 2012, doi: 10.1109/TITS.2012.2225618.

Zhang, Y., & Haghani. A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies: Part B* 58:308-324, https://doi.org/10.1016/j.trc.2015.02.019.

Zhong, Z., Lee, E. E., Nejad, M., & Lee, J. (2020). Influence of CAV clustering strategies on mixed traffic flow characteristics: An analysis of vehicle trajectory data. *Transportation Research Part C: Emerging Technologies* 115:102611.

Table 1. Mobility and event-based features

Feature type*	**Name**	**Description**
Mobility-based Feature	*N_trip*	Number of trips
	Length	Length of the trip (km)
	Duration	Duration of the trip (seconds)
	Speed	Trip speed (km/hr)
Event-based Features (Acceleration based)	*Hacc_count*	Total harsh acceleration event count
	Hacc_max_acc	Maximum acceleration during the harsh acceleration event
	Hacc_avg_acc	Average acceleration during the harsh acceleration event
	Hacc_dur	Duration of the harsh acceleration event
	Hbrk_count	Total harsh braking event count
	Hbrk_max_acc	Maximum acceleration during the harsh braking event
	Hbrk_avg_acc	Average acceleration during the harsh braking event
	Hbrk_dur	Duration of harsh braking the event
	HLacc_count	Total event harsh lateral acceleration count
	HLacc_max_acc	Maximum acceleration during the harsh lateral acceleration event
	HLacc_avg_acc	Average acceleration during the harsh lateral acceleration event
	HLacc_dur	Duration of the harsh lateral acceleration event
Event-based Features (Aggressive Overtaking)	*AO_count*	Total aggressive overtaking event count
	AO_avg_speed	Average speed during the aggressive overtaking event
	AO_max_acc	Maximum acceleration during the aggressive overtaking event
	AO_avg_acc	Average acceleration during the aggressive overtaking event

*All features are computed on the whole period, per day-of-week, and time of the day (morning, afternoon, etc.). Each indicator is aggregated through count, sum, mean, and standard deviation, and is calculated both in total and divided by type of event

Table 2. Socio-economic information of recruited motorcyclists

Characteristics	Units	Frequency (%) (Sample)
Age	1. 18-30 yrs	52
	2. 31-45 yrs	29
	3. 46-60 yrs	13
	4. 60+ yrs	6
Marital Status (*Ms*)	Yes (1) /No (0)	32/68
Education (*Ed*)	1. Below Secondary level	10
	2. Matriculation	28
	3. Intermediate	43
	4. Graduation	19
Household Income (*I*)	1. <20,000 Rs/month	9
	2. 20,000 – 50,000	23
	3. 50,001 – 100,000	29
	4. 100,001 – 150,000	21
	5. 150,001 – 200,000	13
	6. > 200,000	5
Vehicle Availability (*Va*)	1. 1	41
	2. 2	36
	3. 3+	23
Driving Experience (*De*)	No. of years	5.9 (average)
Children (*Chd*)	1. 0	24
	2. 1	31
	3. 2	30
	4. 3+	15
Occupation (*Occ*)	1. Student	18
	2. Full time employee	32
	3. Part time employee	17
	4. Self-business	24
	5. Retired	9
Household Size(*HHS)*	1. 2 or less individuals	24
	2. 3-4 individuals	41
	3. 4-6 individuals	27
	4. 6 and more individuals	8

Table 3. Accident data from the motorcyclists and their statistics

Data type	Units	Aggregate Statistics
Accident type	Fatal	1
	Major Injuries	115
	Minor Injuries/property damage	571
Other vehicles involved	Car	155
	Trucks	136
	Motorcycle	151
	Bicycle	38
	Pedestrian	190
	other	77
No. of people injured	Count	NA
Location and time	Exact location using a map with time	NA
Reported to Police	Yes/No	98/589
Hospital/Clinic visited	Yes/No	201/486
Hospital/Clinic Identification	Name of Hospital/Clinic	NA
Possible cause of Accident	Textual information	NA

Table 4. Model performance statistics for test dataset

Model Performance Statistics	Statistics –test dataset	Statistics -training dataset
Recall-positive	0.83	0.88
Precision-negative	0.98	0.98
F - 1 Score - positive	0.86	0.95

Table 5. Feature importance for the RF model

Features Categories	**Feature**	**Rank**	**Relative Importance (%)**	**Avg. Relative Importance (%)**
Mobility based features	*Average_speed*	18	1.7	1.9%
	Total_distance	7	3.3	
	Total_night_distance	20	1.5	
	Average_morning_speed	22	1.4	
	Std_morning_speed	25	1.1	
	Std_traveltime	23	1.3	
	Average_perday_trips	19	1.6	
	Average_morning_trips	8	3.1	
Event based features (Acceleration based)	*Average_Hbr_acc_max*	16	1.9	2.6%
	Total_Hbr_count	4	4.3	
	Total_Hacc_count	10	2.8	
	Average_morning_Hbr_avg_acc	13	2.3	
	Average_evening_Hbr_avg_acc	5	4.0	
	Average_HLacc_max_acc	21	1.1	
	Std._ Hbr_acc_max	17	1.8	
Event based features (Aggressive Overtaking)	*Total_AO_count*	1	5.5	3.9%
	Std_AO_avg_speed	14	2.2	
	Std_AO_max_acc	11	2.7	
	Average_evening_AO_avg_acc	6	3.7	
	Average_morning_AO_avg_acc	2	5.2	
Motorcyclist's Socio-Economic Features	*Age*	3	4.7	2.7%
	Driving Experience	9	2.9	
	Household Income	24	1.2	
	Education	12	2.2	
	Household size	15	2.5	

Figure captions:

Figure 1: Zig-Zag behaviour or aggressive overtaking behaviour conceptual illustration

Figure 2: The flow diagram of random forest algorithm

Figure 3: Importance of feature categories from RF model

Figure 4: Accident involvement with respect to household size

Figure 5: Accident involvement with respect to household size