Accident Analysis and Prevention 154 (2021) 106094

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw

Muhammad Ijaz [a], Liu lan [a,*], Muhammad Zahid [b], Arshad Jamal [c]

[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan, 610031, China
[b] College of Metropolitan Transportation, Beijing University of Technology, Beijing, 100124, China
[c] Department of Civil and Environmental Engineering, King Fahd University of Petroleum & Minerals, KFUPM, Box 5055, Dhahran, 31261, Saudi Arabia

A R T I C L E I N F O

Keywords:
Three-wheeled motorized rickshaws
Crash severity
Machine learning (ML)
Rawalpindi

A B S T R A C T

Motorcycles and motorcyclists have a variety of attributes that have been found to be a potential contributor to the high liability of vulnerable road users (VRUs). Vulnerable Road Users (VRUs) that include pedestrians, bicyclists, cycle-rickshaw occupants, and motorcyclists constitute by far the highest share of road traffic accidents in developing countries. Motorized three-wheeled Rickshaws (3W-MR) is a popular public transport mode in almost all Pakistani cities and is used primarily for short trips to carry passengers and small-scale goods movement. Despite being an important mode of public transport in the developing world, little work has been done to understand the factors affecting the injury severity of three-wheeled motorized vehicles. Crash injury severity prediction is a promising research target in traffic safety. Traditional statistical models have underlying assumptions and predefined associations, which can yield misleading results if flouted. Machine learning(ML) is an emerging non-parametric method that can effectively capture the non-linear effects of both continuous and discrete variables without prior assumptions and achieve better prediction accuracy. This research analyzed injury severity of three-wheeled motorized rickshaws (3W-MR) using various machine learning-based identification algorithms, i.e., Decision jungle (DJ), Random Forest (RF), and Decision Tree (DT). Three years of crash data (from 2017 to 2019) was collected from Provincial Emergency Response Service RESCUE 1122 for Rawalpindi city, Pakistan. A total of 2,743 3W-MR crashes were reported during the study period that resulted in 258 fatalities. The predictive performance of proposed ML models was assessed using several evaluation metrics such as overall accuracy, macro-average precision, macro-average recall, and geometric means of individual class accuracies. Results revealed that DJ with an overall accuracy of 83.7 % outperformed the DT and RF-based on a stratified 10-fold cross-validation approach. Finally, Spearman correlation analysis showed that factors such as the lighting condition, crashes involving young drivers (aged 20–30 years), facilities with high-speed limits (over 60 mph), weekday, off-peak, and shiny weather conditions were more likely to worsen injury severity of 3W-MR crashes. The outcomes of this study could provide necessary and essential guidance to road safety agencies, particularly in the study area, for proactive implementation of appropriate countermeasures to curb road safety issues pertaining to three-wheeled motorized vehicles.

1. Introduction

Road traffic injuries are the leading cause of injury, and overall the 8th leading cause of all deaths reported worldwide (World Health Organization, 2018). The Global Progress Report on Road Safety points out that the number of annual road traffic deaths has reached 1.35 million in motor vehicle crashes, whereas yearly road traffic injuries (RTIs) are reported to be around as massive as 50 million (World Health Organization, 2018). Every day, over 3,700 people killed worldwide in road traffic crashes involving cars, buses, motorbikes, bicycles, trucks, or pedestrians. More than half of the people killed are pedestrians, motorcyclists, and cyclists that are classified as vulnerable road users (VRUs). According to the World Health Organization (WHO) RTIs are the leading cause of death among children and young adults aged 5–29 years. Studies predict that in the absence of suitable preventive in the near future, motor vehicle fatalities could reach as high as 2 million by

* Corresponding author.
E-mail addresses: m.ijaz58@yahoo.com (M. Ijaz), jianan_l@home.swjtu.edu.cn (L. lan), zahid@emails.bjut.edu.cn (M. Zahid), arshad.jamal@kfupm.edu.sa (A. Jamal).

https://doi.org/10.1016/j.aap.2021.106094
Received 15 July 2020; Received in revised form 3 February 2021; Accepted 15 March 2021
Available online 21 March 2021
0001-4575/

the year 2030 (Organization, 2015). It is reported that around 93 % of the world's traffic deaths result in low-and middle-income nations. The resulting fatality rates are also three-fold higher in low-income countries compared to with-income countries (World Health Organization, 2018). Globally, ambitious goals for road traffic crash prevention have been established in the 2030 plan for sustainable development (Organization, 2015). Projections show that, between 2000 and 2020, road traffic deaths will decline by about 30 % in high-income countries but are anticipated to increase substantially in low-income and middle-income countries (Peden et al., 2004).

Traffic safety is a global issue that is progressing at an alarming rate. Pakistan is a middle-income country with a considerably high rate of road traffic casualties. Pakistan has a high road fatality rate, with approximately 25,781 fatalities per year due to motor vehicle crashes (Organization, 2015; Ahmed et al., 2016). The mean annual fatality rate (deaths/100,000 persons) is 14.3, which is slightly better than the global value of 17.4 (World Health Organization, 2018). It is estimated that two-wheeler motorcycles are involved in approximately 60 % of reported road crashes (Waseem et al., 2019). About 50,000 people have become disabled due to road crashes. The primary sources for crashes in the country include; driver's distractions (use of cell phones and other gadgets), over-speeding, and non-compliance to traffic rules (Ali et al., 2015). Road safety is one of the most neglected social issues in the country. Public transport in the metropolis faces several challenges and their accessibility is limited, encouraging the public to use either personal vehicles (if affordable) or look for other easily available and cheap transport modes.

Motorcycles are a significant concern in transportation safety. From 2002–2017, the number of registered motorcycles in the US rose 100 percent (NHTSA, 2001). In terms of driving style and risk factors for collisions, these types of vehicles vary from regular vehicles. There are advantages associated with motorcycle ridings, such as fuel economy, lower emissions of pollutants, and less parking space requirements (Mannering and Grodsky, 1995). Compared to car crashes, the literature shows that crashes involving motorcycles are potentially more severe (Chen, 2009). These factors fall into four groups, including the attributes of humans, cars, collisions, and roadways. Human attributes include age, gender, residency, and violations of the past. Geometric design, pavement surface types, and weather conditions, roadway, and environmental characteristics. Vehicle attributes, such as the production year and motorcycle size, can also influence the occurrence of motorcycle crashes (Montella et al., 2012; Montella, 2011). Motorized three-wheeled Rickshaws (3W-MR) is a popular public transport mode in almost all Pakistani cities and is used primarily for short trips to carry passengers and small scale goods movement. (Pradhan et al., 2008). Transport networks and road facilities have not met demand because of the rapidly growing population. The use of 3W-MR vehicles as informal transport has increased. However, there is scarce evidence of their safety. Motorized three-wheeled vehicles (mainly including autorickshaws and motorcycle rickshaws) are popular in Asian countries, including Pakistan. The local motorcycle rickshaws (shown in Fig. 1), also commonly called locally as "Ching-chi" are gasoline driven have the capacity to carry a total of six passengers (three on front and back each) and a driver. While autorickshaw (shown in Fig. 1) is normally powered by compressed natural and may accommodate four persons including the driver. The rickshaws were initially brought to the country from China was a hand-pulled vehicle used by merchants to transport goods. In Pakistan motorcycle rickshaws (running on fuel) were introduced in 2007. Autorickshaws were introduced in late 1957 (Khan et al., 2012), powered by a 250 cc, 8 horsepower, 2-stroke, single-cylinder petrol engine. Today several types of 3W-MR may be seen within the cities that are mainly powered by liquefied petroleum gas (LPG), although some are driven by gasoline, as shown in Fig. 1. Their number is increasing every year due to the lack of adequate public transit networks to cope with the growing demand of the metropolis. Travelers prefer to take 3W-MR instead of waiting in long queues for limited transit services. Besides, their travel fare is also low and may be negotiated as well compared to other transport means that make them a very popular mode of commute, particularly for school trips, short-range shopping trips, travel to offices, and several other businesses. However, they are the main source of congestion in urban metropolitan owing to their massive numbers and lack of adequate infrastructure.3W-MR is a form of public transport; hence their growth is desirable from a sustainable transport perspective. Consequently it has been observed that 3W-MR drivers are more frequently involved in traffic violations raising serious road safety issues (Dandona et al., 2005), unlike local buses, as they have no designated stops along the road. Passengers of two-wheeled motorcycles with three wheels combinations (3W-MR) are at greater risk of being thrown out of the vehicle due to a sudden brake application or collision. This throw-off will result in a direct head impact that can lead to head injuries. Attempts may sometimes be made by the occupants to prevent his outstretched hand from falling, which can lead to fractures of the upper limb. In the new rickshaw style, a big improvement is to have a bar between the driver's seat and the passenger seat. By holding this bar, the passenger may gain control of this bar and avoid their fall during crashes or unexpected brake operation.

3W-MR may be classified among potential VRUs group that includes pedestrians and cyclists, motorcyclists, and disabled persons or people with limited mobility and orientation (Ptak, 2019). It is established that VRUs usually have a high risk of crash involvement and more intense crash severity, thereby warranting particular consideration in road safety protection policies (Chandran et al., 2012). VRUs account for nearly 80 % of urban road traffic crash fatalities(Hoque et al., 2008). Although non-motorized three-wheeled vehicles are among the common type of VRUs yet, existing research on injury severity evaluation of 3W-MR is often under-represented. Nearly 75 % of the global 3-Wheeled vehicle population is found in India(Mani and Pant, 2012). Despite their excellent commitment as a mode of transport, they are continuously exposed to significant threats to safety. As a result, speed mitigation policies clearly benefit all VRUs. Efforts to reduce speeding include

Fig. 1. CNG based Autorickshaw (right), and gasoline (petrol) driven motorcycle rickshaw (left).

speed limit setting and enforcement and traffic calming engineering measures (speed bumps, chicanes, roundabouts) (Retting et al., 2003). The latest figures in the US indicate a decrease in accidents for all road users except motorcyclists and bicycles(Administration, N.H.T.S., 2008). In France, where compliance to traffic laws has been enforced strictly since 2002, the proportion of car users among road fatalities dropped by 16 % from 1997 to 2007, while the proportion of VRUs increased by 25 % during the same period. In general, road safety policy is lacking at both national and provincial levels in the country, and in particular, it has no consideration for 3W-MR and other VRUs groups which is an extremely worrying situation (Batool et al., 2012). In Pakistan, the vehicle population, particularly the vulnerable means of transport (particularly motorcycles and 3W-MR) has grown disproportionately over the last decade. In 2018, the number of registered vehicles in the country was 23.5 million indicating an increase of 9.6 % from the preceding year(PBS, 2017). Three-Wheeler data was reported at 1, 044.000 units in Nov 2018(PAMA 2020 PAMA, 2020). This record decrease from the previous number of 1,309.000 Units for Oct 2018. Annual production of 3W-MR during 2019/2020 was over 5,000 indicating a slight decrease from the previous year (Anon., 2021).

A critical bibliographic analysis of the existing literature has shown that three-wheeler crash injury severity is rarely addressed from a global perspective. There are various published studies on road traffic injuries related to motorcycles, but nearly all of them focus on injury patterns and two-wheeled motorcycle collision characteristics. The current study fills this research gap by adopting different machine learning identification algorithms for injury severity analysis of crashes involving 3W-MR in the city of Rawalpindi, Pakistan. Proposed machine learning methods, i.e., Decision Tree (DT), Random Forest (RF), and Decision Jungle (DG) are compared based on several evaluation metrics such as F-1 score (overall prediction accuracies), macro-average precision, macro-average recall, etc. Previous studies for crash severity analysis have mostly reported overall prediction accuracies of method ignoring the model bias toward a specific class. In this study, in addition to the model's overall accuracy, we have shown algorithms' predictive performance for individual severity groups. Finally, Feature importance analysis via Random Forest is carried out to evaluate and define primary risk factors in 3W-MR crashes for current study area. To the best of our knowledge, this is the first study that has investigated the risk factors associated with 3W-Mr crash severity using a machine learning framework. The outcomes of this study could assist in better understanding the causation of 3W-MR crashes, which will have important implications for the identification of targeted prevention measures.

2. Past studies reviews

Predicting injury severity for the individual road user group is an important aspect of road safety research. Crash injury results from a complex interaction among several factors like driver attributes, vehicle conditions, roadway characteristics, features of the built environment, weather conditions, and so forth (Jamal et al., 2020). Previous studies have mostly focused on crash severity analysis of two-wheelers (motorbike and motorcycle) using conventional statistical regression modeling (Truong et al., 2020; Murphy and Morris, 2020; Halbersberg and Lerner, 2019; Montella et al., 2020; Fitzharris et al., 2009; Chang et al., 2019; Vlahogianni et al., 2012). However, it is well known that such methods have several underlying assumptions and predefined associations and are unable to capture latent heterogeneity among predictor variables, also suffer from low prediction accuracies. Very few studies have investigated the injury and vulnerability patterns of 3W-MR (Amarasingha, 2016; Schmucker et al., 2011; de Silva et al., 2014) also referred to as autorickshaws in some published literature. Even though three-wheeler vehicles are the most popular mode of transportation in developing countries, particularly in South Asia as well as in the Far East and African countries, limited research efforts have been undertaken regarding their in-depth crash analysis. The nature of Injuries suffered in three-wheeler crashes, and contributing factors are not well known. Existing literature on three-wheeler crash injury severity is relatively scarce. Particularly in Pakistan, no study has been conducted that has focused on injury severity of crashes involving 3W-MR. In Asian countries, 3W-MR has been considered the hot topic for analysis of three-wheeled vehicles, and nearly 2 million vehicles are reported to be on Indian roads carrying approximately 6–8 billion passenger-km / year (Rajvanshi, 2000). Muzammal et al. revealed in their studies that overloaded and overcrowded Three-wheeled vehicles had a greater impact on crash severity resulted from a 3W-MR crash (Bandara et al., 2019). Bandara et al. revealed in their study that the peak age of the injured patient was between 21–30 years in the 3W-MR crash (Muzzammil et al., 2017). Morency et al. examined the impact of vehicle configuration on three-wheeler stability and found that there is a greater tendency for the wheel lift-off when a three-wheeler hits the road bumps at high speed(Morency et al., 2012). Rehman et al. reported in their study that if the proportion of three-wheelers in the mixed traffic stream is more than 50 percent at the signalized urban intersection, traffic throughput is significantly decreased(RAHMAN et al., 2004). Meena et al. suggested in their work that the most common cause of injury was a collision with a moving vehicle(56 % patients) followed by a fall from a three-wheeler(Meena et al., 2014). Amarasingha conducted a detailed study for the characterization of crashes involving three-wheelers and found that most of the three-wheeler crashes were reported to have happened during the daytime and favorable weather conditions(Amarasingha, 2015). Raman et al. observed that three-wheeled vehicles are more prone to overturning when attempting to avoid obstacles and easy turns(Raman et al., 1995). Tay et al. consider collision type to be a significant indicator impacting Three-Wheeled vehicle crash severity(Tay and Rifaat, 2007). An overturned crash was described as the most important accident element adding to the seriousness of the Three-Wheeled vehicle crash. Kelarestaghi et al. developed a binary logistic model to investigate the predominant factors causing three-wheeled vehicle crashes using Spearman correlation analysis(Amarasingha, 2016). The authors found that factors such as weekdays, daytime driving, travel along rural highways and under fair weather conditions tend to increase the resulting injury severity from Three-wheeled vehicles. Kumarage and, al. revealed in their study that the age of three-wheeler drivers is a significant factor associated with three-wheeled vehicle accident severity (Kumarage et al., 2010). Rahul Goel developed an ecological regression model at the state level to understand the relationship between road deaths and commute distance traveled by different modes in India (Goel, 2018). Model estimation analysis showed that walking, cycling, and intermediate public transport (IPT) modes such as three-wheeled auto-rickshaw (tuk-tuk) are associated with a lower risk of road deaths in states in the country. Authors argued IPTs riders are usually more safe compare to two-wheeled motorcycles. In another study, Rahul Goel et al. adopted a Bayesian hierarchical approach with a Poisson-lognormal regression model and presented a spatial analysis for road fatalities of different VRUs groups with wards as a real unit (Goel et al., 2018). Authors found that crashes involving mixed VRUs groups (Motorized two-wheeler and Three-wheeler vehicle) with heavyweight vehicle weight frequently led to greater injury risk. High speed was identified as another risk factor that intensified the crash severity of occupants on two and three-wheeled motorized vehicles. Hyder et al. conducted a national study and concluded that only 37 individuals per 10,000 registered vehicles were injured annually(Hyder et al., 2000). A study conducted by Ghaffar et al. found that approximately 1,500 road crash injuries per 100,000 population occur annually(Ghaffar et al., 2004). Similarly, Fatmi et al. have established that approximately 1,700 individuals per 100,000 population are injured annually by road traffic injuries in Pakistan(Fatmi et al., 2007). In recent years many researchers have investigated the crash severity identification and prediction performance by comparing different machine learning, data mining, and statistical methods. Zhang, Jian, et al. calculated the correct prediction rate and overall correct prediction rate for each crash severity level(Zhang et al., 2018). Results have shown that machine learning methods have better predictive precision than statistical methods, although they have suffered from an over-fitting problem(Zahid et al., 2020a, b). The RF approach had the highest prediction in both the overall and severe crashes, while the statistical method was the worst. Theofilatos, A., Chen, C., & Antoniou, C. (2019) have compared the machine learning and deep learning models to examine the impact of real-time traffic and weather parameters on crash occurrence on freeways(Theofilatos et al., 2019). The result showed that deep learning models seems to be more appropriate, because it outperformed all the other candidate models. More specifically, the dee model managed to achieve a balanced performance among all metrics compared with other models (total accuracy = 68.95 %, sensitivity = 0.521, specificity = 0.77, AUC = 0.641). Iranitalab, Amirfarrokh, and Aemal Khattak investigated and compared the performance of four statistical and machine learning methods including Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF) (Iranitalab and Khattak, 2017). The result revealed that NNC had the best prediction performance in overall and in more severe crashes. RF and SVM had the next two enough performances and MNL were the weakest method. There are various researcher who investigated and predicted the traffic accidents with data mining techniques which are only related to motorcycles are listed in Table 1. There are various risk factors indicated in the literature compatible with our analysis results. Rider attributes such as riders 20−30 years, facilities with high speed limits e.g. above 60kmph and temporal characteristics such as weekday, off peak and shiny weather condition are the most exposed to risk of injury to 3W-MR crashes. These findings somewhat support other research efforts reported previously(Cunto and Ferreira, 2017; De Lapparent, 2006; Jamson and Chorlton, 2009; Ospina-Mateus et al., 2019; Lin and Kraus, 2008; Wang and Kim, 2019; Zhang et al., 2020; Mokhtarimousavi et al., 2019).In addition, temporal instability is likely to exist for a variety of basic behavioral reasons, and this temporal instability is supported by the results of many recent accident-data analyses(Mannering, 2018; Malyshkina and Mannering, 2009).

3. Study area and data description

The city of Rawalpindi, usually known as "Pindi," was chosen as the study area for this study. It is in the northern part of Pakistan and is adjacent to the capital Islamabad. The two cities are also generally known as "twin cities" due to substantial social and economic ties between them. According to Pakistan bureau statistics, the city has a population of 5.40million inhabitants and covers an area of approximately 5.286 km2. Rawalpindi faces severe seasonal fluctuations in monthly rainfall. Rawalpindi lies at 497 m above sea level. The climate in Rawalpindi is warm and temperate. According to Köppen and Geiger, the climate is classified as Cwa. The average annual temperature in Rawalpindi is 21.5 °C 70.8 °F. Annual precipitation is 941 mm 37.0 in.. 3W-MR crash data (from 2017 to 2019) utilized in this research was collected from the Rawalpindi office Rescue 1122. The Punjab Emergency Service (Rescue 1122) is Pakistan's leading ambulance relief agency with services in all 36 Punjab districts and continues to offer operational assistance to other provinces. Rescue 1122 keeps track of all casualties in this region, including road accidents. The agency uses a two-page incident Statement form to report the incident's precise crash details. Crash reports include the victim's demographic information such as age, class, location period, crash date, and type of vehicle involved in an accident. During the study period, a total of 2,743 road crashes involving 3W-MR (including autorickshaw and motorcycle rickshaw) were reported. Reported crash injuries are divided into four severity levels; No injury, Minor Injury, Severe Injury, and Fatal injury. Among the 2,743 3W-MR 84 (3.06 %) had no injuries, 1551 (56.56 %) had minor injuries, 849 (30.96 %) had serious injuries such as neck fracture, brain injury, and 258 (9.41 %) had fatal injuries as shown in Table 2. Climate and weather data were obtained from the Pakistan metrological department Islamabad. Information about round inventory and geometric features of the roadway, such as numbers of lanes posted speed limits, were also collected from the Rawalpindi Development Authority (RDA).

Approximately three-quarters of 3W-MR crashes happened during the daytime. Detailed descriptive analysis revealed that one-third (34.39 %) of 3W-MR crashes occurred during the summer months. Crash

Table 1
Summary of crash injury severity identification using data mining techniques.

Data Description	Number of injury severity classes	Percent distribution of severity classes	Classification and Identification Algorithm used	Model prediction accuracy (%)	Refs.
5-years of the data (2011−2015) from Ghana (N = 8516)	Damage/ Minor Injury/ Major Injury/ Fatal	5.6/29.4/42.1/22.9	Random Forest	74.0 %	(Wahab and Jiang, 2019)
6-years of data (2008−2013)from Abu Dhabi (N = 5973)	Minor/Moderate/Severe/Death	59/31/7/3	Multilayer Perceptron	77.40 %	(Kunt et al., 2011)
2-years of the data (2005−2006) from Taiwan (N = 1620)	No Injury/Injury/Fatal Injury	60.2/35.0/4.80	Classification and Regression Tree	67.7 %	(Chang and Chien, 2013)
3-years of data (2006−2008) data from Itlay (N = 254,575)	Fatal/Injury	1.8/98.2	Classification trees and Rules discovery	–	(Montella et al., 2012)
11-years of data (2006−2016) data from Wyoming (N = 2430)	Severe and fatal/PDO	34/66	Recurrent Neural Network	78 %	(Rezapour et al., 2020)
5-years of the data (2011−2015) from Ghana (N = 8516)	Damage/ Minor Injury/ Major Injury/ Fatal	5.6/29.4/42.1/22.9	Classification and Regression Tree	73.81	(Wahab and Jiang, 2020)
5-years of the crash data (2011−2015) from Washington (N = 308,641)	PDO/Possible Injury/Evident Injury/Severe and Fatal Injury	82.6/13.1/3.6/0.8	Gradient Boost	82.4 %	(Jiang et al., 2020)
5-years of the crash data (2010−2014) from Louisiana (N = 6,853)	Possible complaint/Non incapacitating/No injury/ Incapacitating/fatal	33.8/33.3/21.3/7.0/ 4.6	Deep Scooter	92.0 %	(Das et al., 2018)
5-years of the crash data (2010−2014) from Uttarakhand, India (N = 14,709)	Killed or severe injury/slight injury	34/66	Decision Tree	90 %	(Kumar and Toshniwal, 2017)
3-years of the crash data (2017−2019) from Saudi Arabia, (N = 12,566)	Fatal/non-fatal	7/93	Feed-Forward Neural Network	77.5 %	(Jamal and Umer, 2020)

Table 2
Summary Statistics of Key Variables.

Variable categories	Variables	Percentage %	Number of crashes				
			No Injury	Minor	Severe	Fatal	Total
Dependent variables							
Level of Injury severity	No Injury	3.06					84
	Minor	56.56					1551
	Severe	30.96					849
	Fatal	9.41					258
Independent variables							
Year	2017	28.30	14	408	278	76	776
	2018	31.80	8	519	262	83	872
	2019	39.90	62	624	309	99	1094
Time	AM	32.24	36	524	251	73	884
	PM	67.76	48	1027	598	185	1858
Gender	Male	80.27	84	1541	373	203	2201
	Female	19.73	–	10	476	55	541
Rider age	< 20yrs,	20.10	11	280	215	45	551
	20−30yrs,	35.26	43	611	261	52	967
	>30−40,	19.83	15	334	141	54	544
	>40−50,	13.63	7	198	128	41	374
	Above 50	11.18	8	128	104	66	306
Day of the week	Sun	15.68	16	228	157	29	430
	Mon	13.75	10	202	126	39	377
	Tue	13.46	8	229	94	38	369
	Wed	12.40	9	204	97	30	340
	Thu	14.88	6	232	122	48	408
	Fri	15.35	21	242	121	39	421
	Sat	14.48	14	214	134	35	397
Nature of Weekday	Weekday	69.84	54	1109	558	194	1915
	Weekend	30.16	30	442	291	64	827
Season of the Year	Winter	29.29	29	460	226	88	803
	Spring	19.07	27	284	160	52	523
	Summer	34.39	22	533	312	76	943
	Autumn	17.25	6	274	151	42	473
Weather Condition	Shiny	67.03	54	1053	559	172	1838
	Rainy	22.72	16	342	203	62	623
	Cloudy	10.25	14	156	87	24	281
Lightening condition	Day	73.12	70	1097	643	195	2005
	Night	26.88	14	454	206	63	737
Period of Hour	Peak	46.17	42	703	378	143	1266
	Off Peak	53.83	42	848	471	115	1476
Posted Speed Limits	≥70 kmph	11.41	11	188	94	20	313
	≥60kmph<	68.67	65	1053	577	188	1883
	70kmph below 60kmh	19.20	8	310	178	50	546
No of Lanes	2	19.91	8	310	178	50	546
	3	68.67	65	1053	577	188	1883
	4	11.12	11	180	94	24	305
	5	.29	–	8	–	–	8
Accident Cause	Over speeding	81.22	28	1441	733	25	2227
	Distraction	13.38	1	71	65	230	367
	U Turn	2.92	28	19	30	3	80
	Wrong Turn	1.46	27	20	21	–	68
Crash Characteristics	Heavy Vehicle	3.13	1	50	35	–	86
	Motorbike	41.06	48	811	257	10	1126
	Passenger Car						
	Roadside Fixed	11.52	9	319	199	5	532
	Object	1.13	1	86	38	2	127
	Pedestrian	9.08	–	12	15	222	249
	3Ws Overturned	20.64	22	237	28	18	566
	Two Rickshaw collided	13.44	3	36	16	1	56

proportion during the weekdays (69.86 %), off-peak hours (53.83 %), and under sparkling weather conditions (67.03 %) were considerably high. The highest of 3W-MR were caused by over-speeding (81.22 %) followed by distraction (13.38 %), U-turn (2.92 %), and wrong turn (1.46 %). The ratio of Motorized 3W-MR crashes was significantly higher for the year 2019(39.90 %) than the previous two years. Drivers or passengers aged 20−30 also have the greatest percentage of contribution to crashes (35.26 %). Most of the crashes happened on the three-lanes highway (68.67 %) with a speed limit of more than 50kmp and less than 70kmp (19.20 %). It has been observed that the accident between two-wheeler motorized motorbike and 3W-MR contributes the largest number of collisions (41.06 %) accompanied by the flipped Three-Wheeled vehicles (20.64 %) since its 3W-MR and being a lightweight and flexible vehicle, this implies substantial higher overturning risk.

4. Pre-processing

3W-MR crash data was collected in the form of an Excel spreadsheet. Pre-processing on the data collection was carried out prior to the implementation of machine learning techniques. Incomplete data, such as data that lack attribute values, missing values in the information were removed from the data set. Data that were incompatible with the names or codes of other data recorded in the data collection were filtered out of the records. Table 2 presents the concluding list of attributes and their descriptions. Current 3W-MR crashes data covers all 2,743 crash records with 16 attributes.

5. Materials and methods

This section reviews the machine learning classification algorithms and discusses their application for 3W-MR crash severity identification. Classifiers are supervised machine learning algorithms that are used in classifying datasets and able to produce promising results due to their multi-dimensional data processing capability, flexibility in implementation, versatility, and superior predictive capabilities. The target variable was 3W-MR crash injury severity, a variable with four possible outcomes (fatal, minor injury, severe injury, and no injury), whereas the independent variables include the temporal features, roadway inventory data, weather conditions, driver sociodemographic attributes, vehicle characteristics, traffic characteristics shown in Table 2. Three crash severity identification algorithms, i.e., Decision Jungle, Random Forest, and Decision Tree, were implemented using stratified 10-fold cross-validation. Moreover, a detailed spearman's correlation analysis features importance score analysis was also made to investigate the correlation among different attributes. In addition, feature importance score and decision rules were generated via the orange data mining toolbox. The detailed methodology is further discussed in the following passages.

5.1. Decision jungle

Decision jungles are the latest addition to decision forests. DAGs are comprised of a set of decision-making acyclic graphs (DAGs). Unlike standard decision trees, the DAG in the decision jungle enables different paths from the root to the leaf. A DAG decision has a reduced memory footprint and provides superior efficiency than a decision tree. Decision jungles are deemed as non-parametric models that provide integrated feature selection, classification, and are robust in the presence of noisy features. DAGs have the same structure as decision trees, except that the nodes have multiple parents (World Health Organization, 2018). Considerin202g the nodes set at two consecutive levels of DAGs, Fig. 2 shows the nodes set consist of child nodes N_c and parent nodes N_p. Let θ_i represents parameters of the split function f for the parent node $i \in N_p$. Si denotes the categorized training samples (x, y), such that it reaches the node i, and can be calculated the set of samples from node i, which travels through its left or right branches. Given θ_i and S_i, the left and right are computed by $S_i^L(\theta_i) = \{(x, y) \in S_i | f(\theta_i, x) \leq 0\}$ and, $S_i^R(\theta_i) = S_i \setminus S_i^L(\theta_i)$ respectively. $l_i \in N_c$ denotes left outward edge from parental node $i \in N_p$ to a child node and $r_i \in N_c$ denotes the right outward edge. Henceforth, the number of samples reaching any child node $k \in N_c$ is given as:

$$S_k(\{\theta_i\}, \{l_i\}, \{r_i\}) = \left[\bigcup_{i \in N_p s.t. l_i = k} S_i^L(\theta_i) \right] \cup \left[\bigcup_{i \in N_p s.t. r_i = k} S_i^R(\theta_i) \right] \quad (1)$$

The model depends on several parameters, which are essential for the efficacy of the model. In order to find improved results and high accuracy, we used the random grid as a hyperparameters optimization for Decision Jungle. The range of best combination for hyperparameters optimization used crash injury severity identification is given in Table 3.

The above listed optimized hyperparameters were achieved using stratified 10-fold cross-validation. For selecting the best hyperparameters of the model, the number of iterations performed was 25. In our case study, the optimized parameters of the experiment demonstrated better performance for the crash injury severity identification and classification. The optimized parameters for Decisions Jungles in Table 3 are further depicted in Fig. 3.

5.2. Random forest

Random Forest is a supervised algorithm for classification. This approach was proposed by (Breiman, 2001), where not only subsets of data are reproduced, but subsets of input variables are also randomly chosen. This describes the most effective and sophisticated tree setting techniques within the traditional or most common strategy. RFs are a combination of decision trees since each tree relies on the values of an individually sampled random vector and with the same distribution for all forest trees. In RF, each node is divided using the best of a subset of randomly selected predictors at that node. Compared to many other widely known classifiers, this method has performed very well and is robust against overfitting. Implementation of this method requires the determination of the parameters of the models, such as the number of trees to grow and the number of variables sampled randomly as candidates at each split. Moreover, selecting a higher value of layers to split the tree nodes helps increase the correlation between trees that yield identical results while voting. The general formula of the Gini index is defined as follows,

$$Gini\ Impurity(p) = 1 - \sum_{i=1}^{N} p_i^2 \quad (2)$$

Where p is the dataset, N is the number of classes i.e. (no injury, minor

Table 3
Range of hyperparameters for 3W-MR injury severity identification.

Parameters description	Values/Range
Resampling method	Bagging
Maximum depth of the decision (DAGs)	{5–500}
Number of optimization steps per decision DAGs layer	{4198–29594}
Maximum width of the decision DAGs	{10–7377}
Number of decisions DAGs	{1–900}

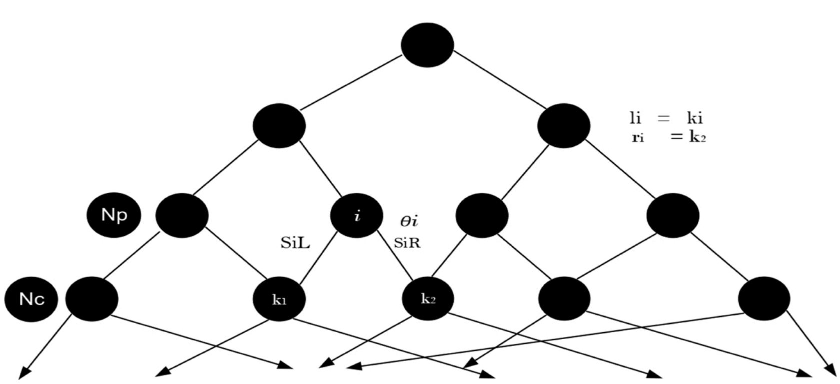


Fig. 2. Schematic for decision jungle.

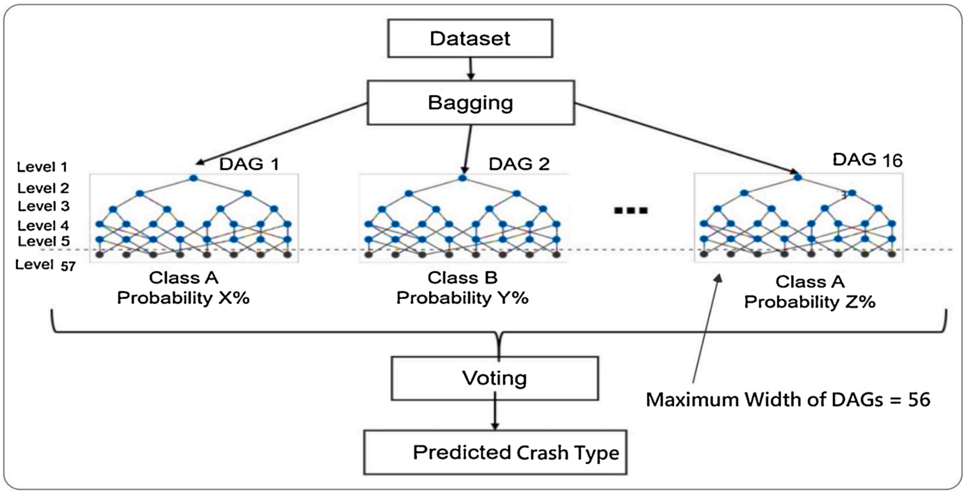


Fig. 3. Structure of DAGs.

injury, severe injury and fatal injury), p_i is the *i* class frequency in *p* the dataset. In addition, the ability to generate rules from the trained decision tree model serves as an advantage for accident severity **identification** and classification.

5.3. Decision tree

The decision tree technique, classified as a regression and classification tree (CART) (Breiman et al., 1984), is a common data mining method used in road safety studies(Chang and Chen, 2005). In the current study, the target variable is the crash injury severity of 3W-MR, which is a discrete variable. Hence, a classification tree is developed. The CART model consists of a hierarchy of univariate binary decisions such as an inverted tree growing from top to bottom. An internal node in the tree defines a binary check on a single variable, each branch is a test outcome and each leaf node, and each leaf node represents a class label or class distribution. A CART chooses the most appropriate vector to split the data into two classes at the root node, splitting the data into two disjoint divisions such that the class names in each division are as homogeneous as possible.

If dataset *T* contains examples from *n* classes i.e. (no injury, minor injury, severe injury and fatal injury), the Gini index (*gini*(*T*)) is defined as follows, where *pj* is the relative frequency of class *j* in dataset *T,*

$$gini(T) = 1 - \sum j = 1\ to\ n\ pj^2 \tag{3}$$

The Gini Split Index contains examples from n classes i.e. (no injury, minor injury, severe injury and fatal injury) when the dataset T is split into two subsets *T1* and *T2* with sizes *N1* and *N2*, *gini*(T) is set as follows:

$$ginisplit(T) = N1/Ngini(T1) + N2/Ngini(T2) \tag{4}$$

Then CART looks at univariate splits in detail. The smallest *gini* divide (T) attribute is selected to divide the node. CART expands the tree recurrently from a root node and then shrugs the large tree gradually.

6. Performance metrics

To assess the model predictive performance, this study utilized various classification evaluation performance metrics such as: overall accuracy, macro-average recall, macro-average accuracy, and geometric mean. Accuracy is the proportion of the correct sample to the total number of samples and can be calculated from Eq. 5; The macro-average recall is the ratio of recall values of each class and the total number of classes and can be calculated from Eq. 6. Likewise, the macro-average precision is the precision of individual classes to the total number of classes (refer to Eq. 7). The geometric mean is defined as the nth root of the product of all values, where n is the number of values and can be calculated from Eq. 8.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Macro - Average - Recall = \frac{(R_1 + R_2 + R_3 + R_4)}{4} \tag{6}$$

$$Macro - Average - Precision = \frac{(P_1 + P_2 + P_3 + P_4)}{4} \tag{7}$$

$$GeometricMean(G.M) = (x_1, x_2, x_3, ..., x_n)^{1/n} \tag{8}$$

Where *R* is recall, *P* is precision and *n* represent the number of periods

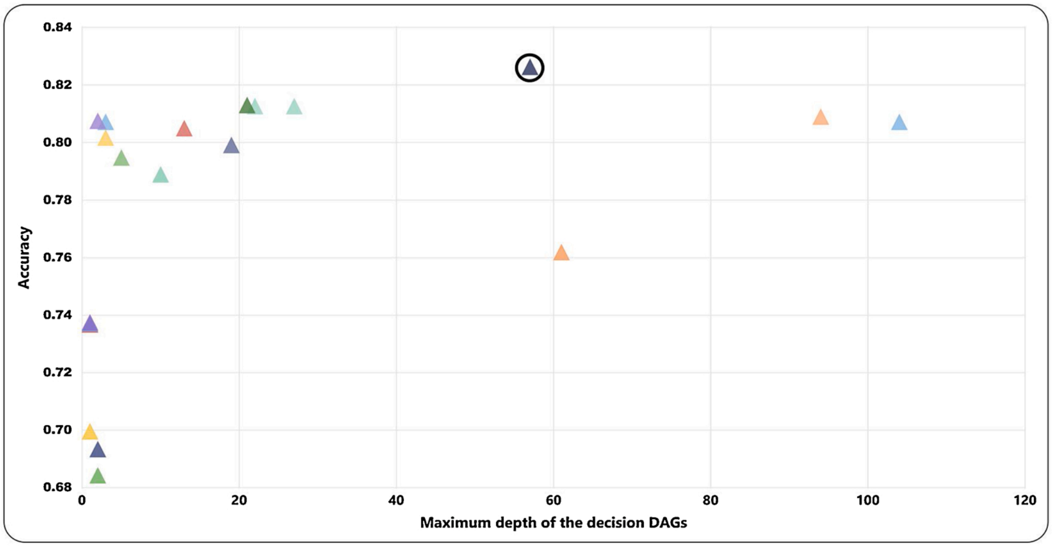
6.1. Stratified 10-fold cross-validation

Stratified K-fold cross-validation is simply the variation of K-fold cross-validation that results in the stratified fold. The data is divided into k folds (k = 10) in k-fold cross-validation. The model is trained on k-1 folds and is held for testing with one-fold. The method is repeated to ensure that each fold of the dataset is preserved. This technique has the advantage over repeated random sub-sampling as all the samples are used for training as well as in the validation, where each sample is used once for the validation. Regarding classification issues, one usually uses stratified k-fold cross-validation under which the folds are chosen such that each fold includes roughly the same proportions of the class labels. In this research, stratified K-fold cross-validation was utilized to solve the problems and bias linked with small and imbalanced datasets as it preserves the percentage of samples for each class or group.

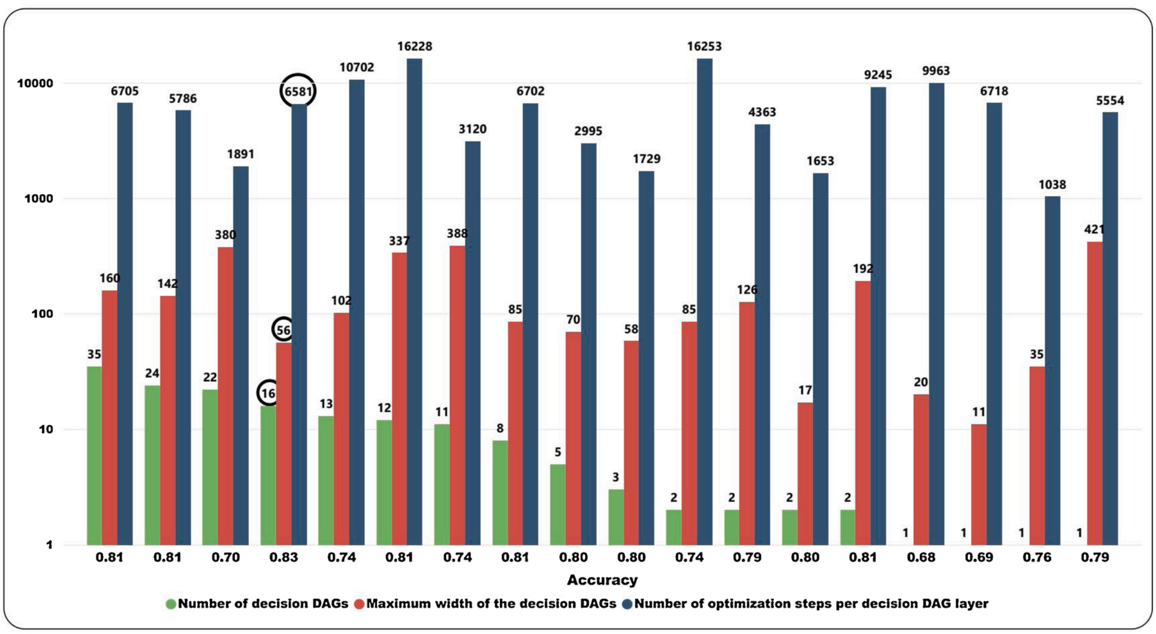
7. Results and discussion

7.1. Model performance evaluation

The aim of building a model for crash injury severity analysis is to explore the relationship between different attributes and crash outcomes. In this study, the tuning parameters in the decision jungle model are the maximum depth of the decision (DAGs), number of decisions DAGs, number of optimization steps per decision DAGs layer, and the maximum width of the decision DAGs'. Fig. 4a shows the impact of the maximum depth of decision DAGs on the overall accuracy of the model. The accuracy was 83.7 % and was achieved when the maximum depth of the decision (DAGs) was 57. Fig. 4b shows the best-optimized values for the other parameters, such as the number of decisions DAGs, number of



(a)



(b)

Fig. 4. Decision Jungle Model a) The impact of maximum depth on accuracy, b) Impact of maximum width of the decision DAGs, number of decisions DAGs on accuracy and number of optimization steps per decision DAGs.

optimization steps per decision DAGs, and maximum width of decision DAGs' were 16, 6581, and 56, respectively. The encircled values in Fig. 4a and 4b show the best-optimized parameter values for applied models for 3W-MR crash severity. The random forest parameters include tree depth, minimum node size, and the number of models. The values of these parameters were 10, 1, and 100 when accuracy and geometric mean were 0.81 and 0.90. Similarly, the decision Tree parameters include the minimum number of records per node and the number of threads. The accuracy and Geometric mean obtained for the decision tree was 76.6 % and 87 % when the values of parameters were 2 and 8, as shown in Fig. 5. The Geo-metric mean for Decision Jungle was 0.93. Fig. 5 shows the performance comparison of applied models for 3W-MR crash severity. In addition, the average accuracy, macro-average precision, and macro-average recall were evaluated in order to assess the performances of different models for 3W-MR crash severity. Table 4 shows the macro-average precision and macro-average recall for different models, as well as the average accuracy for each crash type. The results were attained using stratified 10-fold cross-validation. Moreover, the results showed that the Decision jungle outperformed the Decision Tree and Random Forest. In literature different data mining techniques have been used for 2 wheeler crash injury severity prediction.Al-radaideh and Daoud have applied three different machine learning algorithms such as random forest, SVM and ANN for injury severity prediction. The result showed that Random forest with 80.6 % accuracy outfperformed ANN(61.4 %) and SVM (54.8 %)(Al-Radaideh and Daoud, 2018).Similarly, Kumar, S., & Toshniwal D, analyzed the

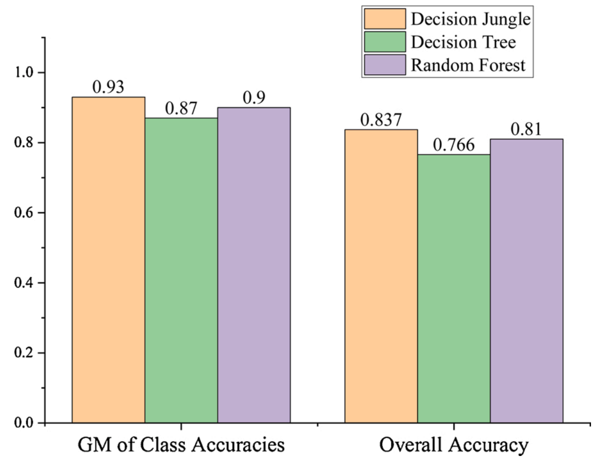


Fig. 5. Overall accuracy, and geometric mean (GM) of class accuracies for different models.

factors that effect the road injury severity using three different machine learning algorithms i.e. decision tree, Naive bayes and SVM. The results revealed that decision tree algorithms classification accuracy was found better than the other two technicques(Kumar and Toshniwal, 2017). Algorithms such as ANNs and SVMs have a good ability to predict and classify the data, they cannot provide proper interpretation and look more like a 'black-box' difficult to understand and provide individualized feedback. Therefore, these algorithms cannot be employed in practice to play a role to decrease and control accidents. In contrast to algorithms such as ANNs and SVM, rule-based algorithms like DTs and random forest have the ability to be interpreted and understood easily.

Fig. 6 shows the receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics for evaluating the predictive performance of the proposed crash identification algorithms. The AUC is the area enclosed by the ROC curve with a maximum value 1, indicating a perfect classification result. Whereas, a value near 0.5 indicates that the algorithm produces absolutely random classification. Literature suggests that a value of AUC above 0.7 is considered acceptable for classification problems. From the results in Fig. 6, It is clear that the DJ model achieves the best performance with an AUC of 0.956. This is followed by that for RF and DT models, with AUCs of 0.874 and 0.829, respectively.

7.2. Decision rules for crash severity identification

Although feature sensitivity analysis is vital for establishing the significance of individual risk factors in crash injury severity investigation, it is well-known that frequency and injury severity of traffic crashes results from multi-facet interactions among predictor variables. Therefore, it is crucial to identify the circumstances that lead to a particular crash injury severity outcome. In this regard, decision rules (DRs) highlighting the risk factors for specific crash severity groups have proven beneficial for policy recommendations and safety improvement. DRs are important because they can capture the latent correlation among the predictor variables in the crash data, which can guide the authorities in better decisions regarding road safety(Hashmienejad and Hasheminejad, 2017). For the current study, DRs were extracted using the RF technique for classifying the circumstances prevalent to individual crash severity groups. To describe the pattern showed in the rules, only a subset of rules showing coherence with descriptive and inferential analysis results are shown in Table 5. As shown in the table, some of the predominant main factors associated with injury severity of crashes involving 3W-MR include road characteristics, environmental characteristics, human attributes, temporal characteristics, and crash

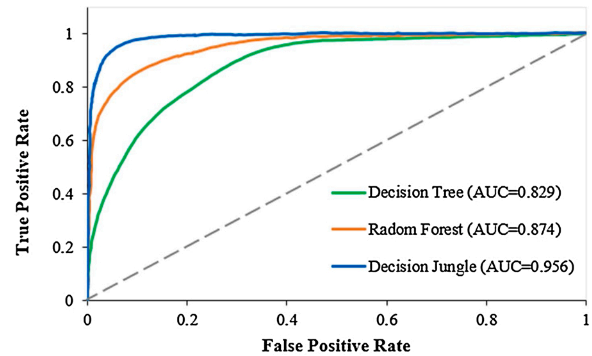


Fig. 6. ROC and AUC metrics for algorithm comparison.

Table 5
The significant decision rules identified by the proposed method.

Num	Rule(IF)	Then	P (%)
1	Crash type = Rollover: Reason = over speeding: time of the day = off peak; Nature of weekday = Weekday	Fatal	76
2	Age = 20−30 years: Crash type = Passenger car; gender = male	Severe	67
3	Crash type= Heavy vehicle: Lighting condition = night: reason = over	Fatal	87
4	speeding	Minor	43
5	Posted Speed limit<60kmph: time = peak time: weather = Shiny	Severe	66
6	Weather = cloudy: Speed limit>70kmph:time = off peak	Fatal	79
7	Reason = Distraction: lightning condition = night	Severe	68
8	Gender = Male; Reason = Wrong way; Weather = Rainy Lighting condition = night; Age= >50 years; crash type = heavy vehicle;	Fatal	83
9	reason = wrong way	Severe	69
10	Age = below 17.5years; crash type: motorbike; reason = over speeding	Minor	54
11	Crash type = Passenger car; reason = faulty vehicle	No Injury	58
12	Gender = Female; Nature of weekday = weekday; crash type = fixed	Fatal	74
13	object/barrier Crash type = pedestrian; reason = over speeding; time = off peak	Minor	66
14	Age = 30−40years;Weather = cloudy; lightning condition = day; Speed limit<60kmph Reason = distraction; no of lanes=>4; lighting condition = day; speed limit>70kmph	Severe	91

Table 4
Summary of classification performance.

• Number of Classes • Class break-down (%, from Class 1 to 4)	Identification algorithms to classify crash injury severity	Classification results from stratified K-Fold cross-validation						
		Average accuracy (%)	Class 1	Class 2	Class 3	Class 4	Macro Average Precision	Macro Average Recall
• 4 • Minor injury (56.56) / Fatal injury (9.41) / Severe injury (30.96) / No injury (3.06)	Random Forest	0.90	0.83	0.97	0.85	0.96	0.75	0.63
	Decision Tree	0.879	0.79	0.955	0.809	0.957	0.68	0.65
	Decision Jungle	0.93	0.85	0.97	0.87	0.98	0.85	0.73

characteristics. Where p is the probability of an outcome to stand true. DRs obtained are graded according to their intensity of crash severity, which can help safety experts to prioritize and concentrate more on fatal or severe injury crashes to mitigate the loss of precious human lives and permanent disabilities. The following points summarize the main conclusions noted from DRs extracted by the proposed method.

- Fatal and severe injury crashes are related to crash characteristics, specifically crash type. Results showed that the probability of severe injury outcomes increases if the vehicle hit is a passenger car, a driver is a male of age below 30 years, and those under cloudly and off-peak conditions.
- It is observed the likelihood of fatal injury is significantly increased for crashes caused due to speeding, driver distractions, and wrong-way driving.
- Extracted rules also revealed that if the gender of the driver is male, driver age is above 50 years, crashes that involve heavy vehicles (trucks) and pedestrians, dark lighting, and rainy weather conditions also amplify the chances of fatal crashes.
- The probability of minor injury increases if the crash type is hit a fixed object/barrier, the crash reason is faulty vehicle component, facilities with a speed limit less 60kmph, and those that occurred in daylight and shiny weather conditions.
- Finally, the extracted rules suggest if the crash type is hit fixed object/barrier, the driver gender is female, and the collisions occurring on weekdays lead to a high probability of no injury crashes.

7.3. Spearman correlation and features importance analysis

To examine the correlation between different predictor variables, the Spearman correlation was performed. The analysis is aimed to better understand if there exists a high correlation between the pair of independent variables based on which some of the variables may be removed from severity modeling experiments to avoid biased predictions. The strength of the relationship between two sets of items, which may be a dependent and independent variable, or even two independent variables (Han et al., 2011). This relationship helps us to identify which independent variables can have a more substantial impact on the dependent variables and therefore leads us to predict more effectively the outcome of the dependent variable. Numerically, this relationship is usually determined by a decimal value known as the coefficient of correlation.

The Spearman Correlation can also be interpreted as to the amount of reciprocal knowledge between two variables, and the value of the Spearman correlation is a measure of the consistency of two sets of graded variables (Myers et al., 2010). The value of the coefficient (C) is between −1 and +1. Values close to +1 show a strong positive correlation, those close to −1 show a strong negative correlation, and those close to 0 show no relationship. Spearman correlation analysis results showed that the variable number of vehicles involved in a crash, patients in an emergency were found to have a highly negative correlation with the target(dependent) variable (3W-MR crash severity), where variables such as year and posted speed limit are the only two factors that were observed to have a strong positive correlation with the target variable, as shown in Fig. 7. To assess the importance of each variable and the frequency of the association between the multiple pairs of variables Kelarestaghi et.al. performed a Spearman correlation test. In his report, he revealed that factors such as adverse weather conditions and young drivers decrease the severity of the crash, while the involvement of motorcycles, pedestrians, unbelted passengers and heavy vehicles significantly raises the risk of a severe accident(Kelarestaghi et al., 2017)

This attribute takes four values as target values. Fig. 8 displays the relative significance rating of each attribute based on a random forest-based feature importance evaluation. Feature importance is a technique that assigns a score to input features based on how important they are to predict a target variable. For current study a random forest-based performance evaluation method was adopted to explore the relative significance of each factor in predicting the crash injury severity. The high value of the evaluator implies a comparatively higher importance of the feature. The findings indicate that the explanatory factors that have a severe effect on the injury severity are weather, type of crash, rider's age, speed limit and lightning condition While the number of lanes, season of the year, gender, number of vehicles involved in the crash and the length of the day are slightly associated with the severity of the injuries. Feature importance results shown herein pertaining to crash injury severity are mostly in agreement with those of previous studies from different regions(Kelarestaghi et al., 2017; Huang et al., 2011; Abdel-Aty, 2003; Fiorentini and Losa, 2020). However, few contributing factors, such as on-site road type injury, road surface conditions, road type, are not commonly investigated in the literature, making them novel and important.

7.4. Proposed mitigation measures

Decision rules for injury severity classifications could provide crucial

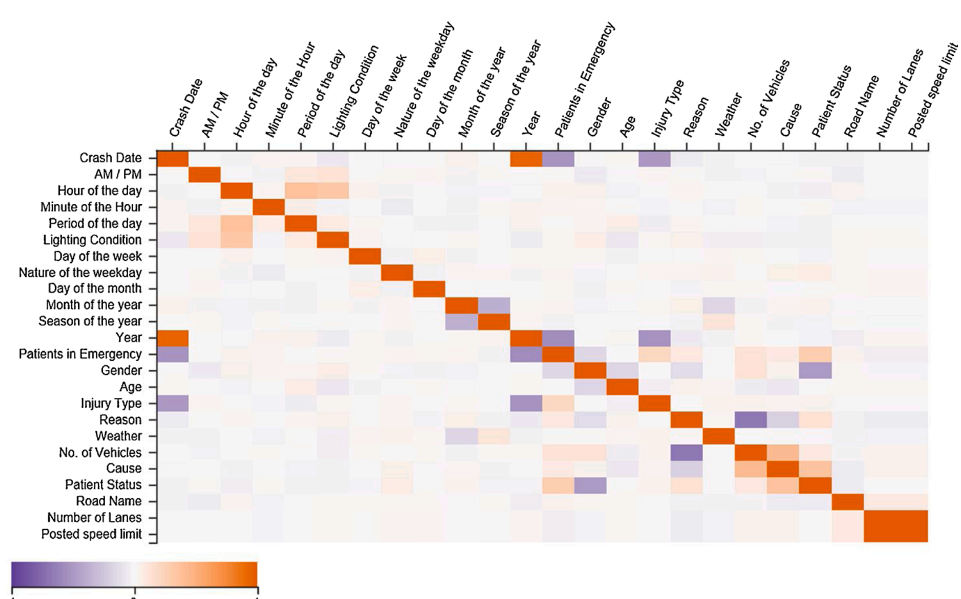


Fig. 7. Spearman's Correlation Matrix.

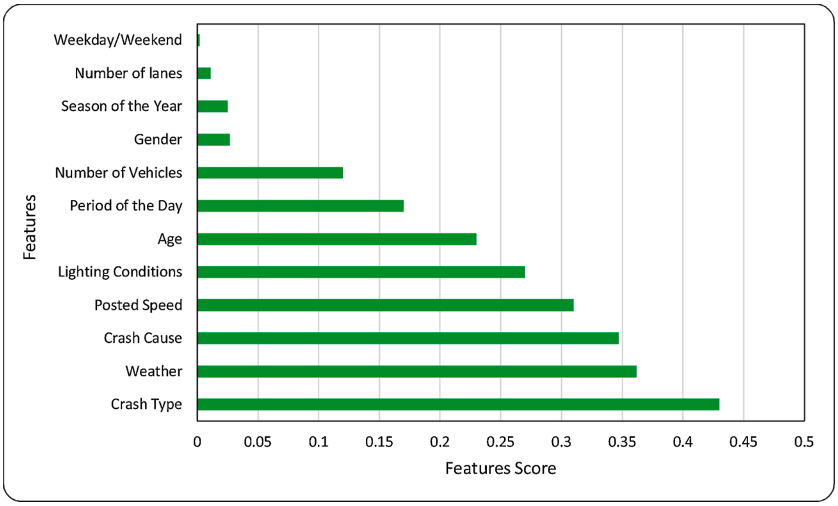


Fig. 8. Features Importance Via Random Forest.

insights for road safety professionals and practitioners to design policy implications and suitable countermeasures to mitigate the occurrence as well as the severity of 3W-MR crashes. Safety improvement decisions based on decision rules are more appropriate and effective instead of considering the factors separately since they incorporate the interplay between them. The rules identified (Table 5) using the 3W-MR crash data are specific for the study area that can guide the agencies to adopt essential mitigation measures proactively. For example, the extracted rules suggested that the likelihood of fatal crashes is associated with factors such as overspeeding (particularly among young male drivers), driver distractions, collisions involving passenger cars, trucks, nighttime travel, off-peak, and adverse weather conditions. To avoid the burden of fatal crashes, several strategies are recommended in this regard. It is recommended to strictly enforce speed regulations by providing speed monitoring cameras and introduce calming traffic measures (speed humps, speed tables, horizontal shifts, etc.) in hazard-prone locations. This can also be supplemented by proper deployment of variable message signs (VMS) for early warnings about congestions, traffic incidents, unexpected delays to reduce driver's stress. Likewise, the provision of VMS could also be very beneficial especially during off-peak and rainy weather conditions. It is also suggested to provide a dedicated lane for two-and-three-wheelers to separate them from main traffic, particularly at crash hotpost locations, which will reduce their direct contact and hence the probability of fatal crashes. It is also advised to increase general public awareness about traffic rules and safety, whereas periodic safety awareness campaigns among young drivers (in particular) through traffic police and volunteer teams may be initiated. Similarly, to restrain the losses of severe and minor injury crashes under the circumstances extracted by DRs (Table 5), some potential strategies that are recommended include improvement of highway delineation, such as raised pavement markers, delineators on horizontal curves, installation of new signs, and replacement of faded ones that can substantially enhance the visual acuity particularly among for elderly drivers during night time. Similarly, provisions of crash cushions at the roadside barrier and fixed object is also recommended that will reduce the intensity of vehicle impact during a crash. It is also witnessed that many underage drivers are operating the 3W-MR that is one of the leading causes of traffic crashes in the area. This practice should be discouraged by legislation and enforcement of child restraint law. At present, helmet use is mandatory only for 2Ws motorbike riders; however, there are no specific regulations for autorickshaws and 3W-MR drivers. The use of helmets among 3W-MR drivers should be legislated and strictly enforced on immediate priority to reduce the burden of such crashes.

8. Conclusions

Road safety is a significant public health issue and attention must be paid to road safety initiatives. Strict adherence to road safety measures prevents casualties from road crashes. Three-wheeled motorized vehicles constitute an essential public transport mode over short trips in developing countries. However, 3W-MR (both motorcycle rickshaw and autorickshaw) may be classified under VRUs that are frequently involved in road traffic crashes resulting in a large number of serious injuries and fatalities. Predicting injury severity of 3W-MR with better accuracy and identification of associated risk factors can guide on sound policy recommendations to improve the safety of both drivers and passengers. This study exclusively focuses on injury severity analysis of 3W-MR crashes using machine learning identification algorithms. Crash dataset was obtained from Rescue 1122 for the city of Rawalpindi, Pakistan. The available data had crash severity defined under four categories, including no injury (3.06 %), minor injury (56.6 %), severe injury (30.46 %), and fatal injury (9.41 %). Severity predictive performance of three machine learning identification algorithms, i.e., DT, RF, and DJ, were compared based on different evaluation metrics such as overall accuracy, macro-average precision, macro-average recall, and the geometric mean of individual severity groups. To enhance prediction accuracy from DJ, hyperparameter optimization with a random grid was accomplished. Study results showed that the DJ with an overall accuracy of 83.7 % outperformed the DT and RF. Models comparison showed increased predictive efficiency and robust performance of the DJ model with reference to other metrics like precision, recall, and the geometric mean of individual crash severity classes. In addition to severity identification, variable importance, and Spearman correlation analysis were performed to explore the key injury severity risk factors in 3W-MR crashes. Results revealed drivers related factors including gender (male), age (20−30 years age group), over speeding were positively associated with injury severity of such crashes. Similarly, some roadway related characteristics (presence three lanes or more, highway with speed limit 60 mph or more), crash temporal variables (driving during the daytime, weekday, and off-peak periods), and shiny weather conditions were also noted to worsen crash severity of 3W-MR. Future studies could seek more advanced techniques such as ensemble and deep learning on other detailed datasets to explore factors contributing to this VRUs group.

Author statement

Muhammad Ijaz: Conceptualization, Methodology, Software, Data curation, preparation, Validation, formal analysis, writing-original draft.

Liu Lan: Conceptualization, Writing-review and editing, Supervision, Funding acquisition.

Muhammad Zahid: Software, Visualization, investigation, Writing-review and editing

Arshad Jamal: formal analysis, Writing-review and editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

We would like to express our gratitude to the National Natural Science Foundation of China, grant number 61873216.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Safety Res. 34, 597–603.
Administration, N.H.T.S, 2008. Traffic Safety Annual Assessment-Highlights. NHTSA Natl. Cent. Stat. Anal.
Ahmed, A., Khan, B.A., Khurshid, M.B., Khan, M.B., Waheed, A., 2016. Estimating national road crash fatalities using aggregate data. Int. J. Inj. Contr. Saf. Promot. 23, 249–254.
Ali, S.A., Kanpurwala, M.A., Aslam, M., Siddiqui, S., Khan, S., Haider, S., Tahir, M., 2015. Association of Motorbike Accidents with Behavior Related Factors in Karachi, Pakistan.
Al-Radaideh, Q.A., Daoud, E.J., 2018. Data mining methods for traffic accident severity prediction. Int. J. Neural Netw. Adv. Appl 5, 1–12.
Amarasingha, N., 2015. Characteristics of three-wheeler crashes. Sri Lanka Inst. Inf. Technol. Sri Lanka.
Amarasingha, N., 2016. Modeling the Severity of Three-wheeler Crashes Using the Data From Sri Lanka.
Anon. Https://www.ceicdata.com/en/pakistan/motor-vehicle-production-pakistan-automotive-manufacturers-association/motor-vehicle-production-pama-motorcycle-sazgar-three-wheeler Monthly Production & Sales of Vehicles Available online: https://www.ceicdata.com/en/pakistan/motor-vehicle-production-pakistan-automotive-manufacturers-association/motor-vehicle-production-pama-motorcycle-sazgar-three-wheeler.
Bandara, K., Vadysinghe, A.N., Edussuriya, D.H., 2019. Medico legal aspects of injuries sustained by occupants of three wheelers in road traffic crashes: a study conducted in a tertiary care hospital in the central province of Sri Lanka. Sri Lanka J. Forensic Med. Sci. Law 10.
Batool, Z., Carsten, O., Jopson, A., 2012. Road safety issues in Pakistan: a case study of Lahore. Transp. Plan. Technol. 35, 31–48. https://doi.org/10.1080/03081060.2012.635415.
Breiman, L., 2001. Random Forests Mach Learn, 45, pp. 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC press. ISBN 0412048418.
Chandran, A., Sousa, T.R.V., Guo, Y., Bishai, D., Pechansky, F., Team, T.V.N.T.E., 2012. Road traffic deaths in Brazil: rising trends in pedestrian and motorcycle occupant deaths. Traffic Inj. Prev. 13, 11–16.
Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36, 365–375.
Chang, L.-Y., Chien, J.-T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51, 17–22.
Chang, F., Xu, P., Zhou, H., Chan, A.H.S., Huang, H., 2019. Investigating injury severities of motorcycle riders: a two-step method integrating latent class cluster analysis and random parameters logit model. Accid. Anal. Prev. 131, 316–326.
Chen, C.-F., 2009. Personality, safety attitudes and risky driving behaviors—evidence from young Taiwanese motorcyclists. Accid. Anal. Prev. 41, 963–968.
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.
Dandona, R., Anil Kumar, G., Dandona, L., 2005. Traffic law enforcement in Hyderabad, India. Int. J. Inj. Contr. Saf. Promot. 12, 167–176.
Das, S., Dutta, A., Dixon, K., Minjares-Kyle, L., Gillette, G., 2018. Using deep learning in severity analysis of at-fault motorcycle rider crashes. Transp. Res. Rec. 2672, 122–134.
De Lapparent, M., 2006. Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. Accid. Anal. Prev. 38, 260–268.
de Silva, M., Nellihala, L.P., Fernando, D., 2014. Pattern of accidents and injuries involving three-wheelers. Ceylon Med. J. 46.
Fatmi, Z., Hadden, W.C., Razzak, J.A., Qureshi, H.I., Hyder, A.A., Pappas, G., 2007. Incidence, patterns and severity of reported unintentional injuries in Pakistan for persons five years and older: results of the National Health Survey of Pakistan 1990–94. BMC Public Health 7, 152.
Fiorentini, N., Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures 5, 61.
Fitzharris, M., Dandona, R., Kumar, G.A., Dandona, L., 2009. Crash characteristics and patterns of injury among hospitalized motorised two-wheeled vehicle users in urban India. BMC Public Health 9, 11.
Ghaffar, A., Hyder, A.A., Masud, T.I., 2004. The burden of road traffic injuries in developing countries: the 1st national injury survey of Pakistan. Public Health 118, 211–217.
Goel, R., 2018. Modelling of road traffic fatalities in India. Accid. Anal. Prev. 112, 105–115.
Goel, R., Jain, P., Tiwari, G., 2018. Correlates of fatality risk of vulnerable road users in Delhi. Accid. Anal. Prev. 111, 86–93.
Halbersberg, D., Lerner, B., 2019. Young driver fatal motorcycle accident analysis by jointly maximizing accuracy and information. Accid. Anal. Prev. 129, 350–361.
Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques. Elsevier. ISBN 0123814804.
Hashmienejad, S.H.-A., Hasheminejad, S.M.H., 2017. Traffic accident severity prediction using a novel multi-objective genetic algorithm. Int. J. crashworthiness 22, 425–440.
Hoque, M.M., Mahmud, S.M.S., Qazi, A.S., 2008. Dealing with vulnerable road user (VRU) safety and mobility in urban areas of Bangladesh: a critical sustainable transport development challenge. In: Proceedings of the Conference Proceedings. XIII CODATU, pp. 12–14.
Huang, H., Siddiqui, C., Abdel-Aty, M., 2011. Indexing crash worthiness and crash aggressivity by vehicle type. Accid. Anal. Prev. 43, 1364–1370.
Hyder, A.A., Ghaffar, A., Masood, T.I.M., 2000. Otor vehicle crashes in Pakistan: the emerging epidemic. Inj. Prev. 6, 199–202.
Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.
Jamal, A., Umer, W., 2020. Exploring the injury severity risk factors in fatal crashes with neural network. Int. J. Environ. Res. Public Health 17, 7466.
Jamal, A., Rahman, M.T., Al-Ahmadi, H.M., Mansoor, U., 2020. The dilemma of road safety in the eastern province of Saudi Arabia: consequences and prevention strategies. Int. J. Environ. Res. Public Health 17, 157.
Jamson, S., Chorlton, K., 2009. The changing nature of motorcycling: patterns of use and rider characteristics. Transp. Res. Part F Traffic Psychol. Behav. 12, 335–346.
Jiang, L., Xie, Y., Wen, X., Ren, T., 2020. Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis. J. Transp. Saf. Secur. 1–23.
Kelarestaghi, K.B., Zhang, W., Wang, Y., Xiao, L., Hancock, K., Heaslip, K.P.I., 2017. Mpacts to crash severity outcome due to adverse weather and other causation factors. Adv. Transp. Stud.: Int. J. 43, 31–42.
Khan, J.H., Hassan, T., Shamsad, S., 2012. Socio-economic profile of cycle rickshaw pullers: a case study. Eur. Sci. J. 8, 310–330.
Kumar, S., Toshniwal, D., 2017. Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India. Eur. Transp. Res. Rev. 9, 24.
Kumarage, A.S., Bandara, M., Munasinghe, D., 2010. Analysis of the economic and social parameters of the Three-Wheeler Taxi service in Sri Lanka. Res. Transp. Econ. 29, 395–400.
Kunt, M.M., Aghayan, I., Noii, N., 2011. Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. Transport 26, 353–366.
Lin, M.-R., Kraus, J.F., 2008. Methodological issues in motorcycle injury epidemiology. Accid. Anal. Prev. 40, 1653–1660.
Malyshkina, N.V., Mannering, F.L., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accid. Anal. Prev. 41, 829–838.
Mani, A., Pant, P., 2012. Review of Literature in India's Urban Auto-rickshaw Sector: A Synthesis of Findings.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.
Mannering, F.L., Grodsky, L.L., 1995. Statistical analysis of motorcyclists' perceived accident risk. Accid. Anal. Prev. 27, 21–31.
Meena, S., Barwar, N., Rastogi, D., Sharma, V., 2014. Injuries associated with cycle rickshaws accidents. J. Emerg. Trauma Shock 7, 73.
Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transp. Res. Rec. 2673, 680–692.
Montella, A., 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accid. Anal. Prev. 43, 1451–1463.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.
Montella, A., de Oña, R., Mauriello, F., Riccardi, M.R., Silvestro, G., 2020. A data mining approach to investigate patterns of powered two-wheeler crashes in Spain. Accid. Anal. Prev. 134, 105251.
Morency, P., Gauvin, L., Plante, C., Fournier, M., Morency, C., 2012. Neighborhood social inequalities in road traffic injuries: the influence of traffic volume and road design. Am. J. Public Health 102, 1112–1119.
Murphy, P., Morris, A., 2020. Quantifying accident risk and severity due to speed from the reaction point to the critical conflict in fatal motorcycle accidents. Accid. Anal. Prev. 141, 105548.
Muzzammil, M., Minhas, M.S., Effendi, J., Jahanzeb, S., Mughal, A., Qadir, A., 2017. Qing-qi rickshaw: a boon or bane for public transportation? A study of road traffic injury patterns involving qing-qi rickshaws in Karachi Pakistan. J. Ayub Med. Coll. Abbottabad 29, 289–292.
Myers, J.L., Well, A., Lorch, R.F., 2010. Research Design and Statistical Analysis. Routledge. ISBN 0805864318.
NHTSA, U.S., 2001. Traffic Safety Facts 2000.

Organization, W.H., 2015. Global Status Report on Road Safety 2015. World Health Organization. ISBN 9241565063.
Ospina-Mateus, H., Jiménez, L.A.Q., Lopez-Valdes, F.J., Salas-Navarro, K., 2019. Bibliometric analysis in motorcycle accident research: a global overview. Scientometrics 121, 793–815.
PAMA 2020 PAMA, 2020. Production & Sale of Vehicles Pakistan Automotive Manufacturing Association. Available online. https://www.pama.org.pk/statistical-information/sales-production/monthly-sales-production.
PBS, 2017. Pakistan Statistical Year Book 2017. Available online:. Pakistan Bureau of statistics https://www.pbs.gov.pk/content/pakistan-statistical-year-book-2017.
Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., Mathers, C.D., 2004. World Report on Road Traffic Injury Prevention.
Pradhan, C.K., Thakur, S., Mukherjee, A.K., Roychowdhury, A., 2008. Energy expenditure of cycle rickshaw pullers in different places in India. Ergonomics 51, 1407–1417.
Ptak, M., 2019. Method to assess and enhance vulnerable road user safety during impact loading. Appl. Sci. Basel (Basel) 9. https://doi.org/10.3390/app9051000.
RAHMAN, M.M., OKURA, I., NAKAMURA, F., 2004. Effects of rickshaws and auto-rickshaws on the capacity of urban signalized intersections. IATSS Res. 28, 26–33.
Rajvanshi, A.K.C., 2000. Ycle rickshaws as a sustainable transport system for developing countries. Hum. Power, Tech. J. IHPVA 15–18.
Raman, A., Rao, J.S., Kale, S.R., 1995. Overturning stability of three wheeled motorized vehicles. Veh. Syst. Dyn. 24, 123–144.
Retting, R.A., Ferguson, S.A., McCartt, A.T.A., 2003. Review of evidence-based traffic engineering measures designed to reduce pedestrian–motor vehicle crashes. Am. J. Public Health 93, 1456–1463.
Rezapour, M., Nazneen, S., Ksaibati, K., 2020. Application of deep learning techniques in predicting motorcycle crash severity. Mater. Sci. Eng. R Rep., e12175
Schmucker, U., Dandona, R., Kumar, G.A., Dandona, L., 2011. Crashes involving motorised rickshaws in urban India: characteristics and injury patterns. Injury 42, 104–111.
Tay, R., Rifaat, S.M., 2007. Factors contributing to the severity of intersection crashes. J. Adv. Transp. 41, 245–265.
Theofilatos, A., Chen, C., Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. Transp. Res. Rec. 2673, 169–178.
Truong, L.T., Nguyen, H.T.T., Tay, R., 2020. A random parameter logistic model of fatigue-related motorcycle crash involvement in Hanoi, Vietnam. Accid. Anal. Prev. 144, 105627.
Vlahogianni, E.I., Yannis, G., Golias, J.C.O., 2012. Verview of critical risk factors in Power-Two-Wheeler safety. Accid. Anal. Prev. 49, 12–22.
Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS One 14, e0214966.
Wahab, L., Jiang, H., 2020. Severity prediction of motorcycle crashes with machine learning methods. Int. J. crashworthiness 25, 485–492.
Wang, X., Kim, S.H., 2019. Prediction and factor identification for crash severity: comparison of discrete choice and tree-based models. Transp. Res. Rec. 2673, 640–653.
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. Accid. Anal. Prev. 123, 12–19.
World Health Organization, 2018. World Health Organization Global Status Report on Road Safety. Available online. https://www.who.int/publications/i/item/global-status-report-on-road-safety-2018.
Zahid, M., Chen, Y., Jamal, A., Al-Ofi, K.A., Al-Ahmadi, H.M., 2020a. Adopting machine learning and spatial analysis techniques for driver risk assessment: Insights from a case study. Int. J. Environ. Res. Public Health 17, 5193.
Zahid, M., Chen, Y., Khan, S., Jamal, A., Ijaz, M., Ahmed, T., 2020b. Predicting risky and aggressive driving behavior among taxi drivers: do spatio-temporal attributes matter? Int. J. Environ. Res. Public Health 17. https://doi.org/10.3390/ijerph17113937.
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.
Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., Zhou, B., 2020. A crash severity prediction method based on improved neural network and factor Analysis. Discret. Dyn. Nat. Soc. 2020.