Accident Analysis and Prevention 192 (2023) 107275

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Injury severity prediction of cyclist crashes using random forests and random parameters logit models

Antonella Scarano [a,1,*], Maria Rella Riccardi [a,2], Filomena Mauriello [a,3], Carmelo D'Agostino [b,4], Nicola Pasquino [c,5], Alfonso Montella [a,6]

[a] University of Naples Federico II Department of Civil, Architectural and Environmental Engineering Via Claudio 21, 80125 Naples, Italy
[b] Department of Technology and Society, Faculty of Engineering, LTH Lund University, Lund, Sweden
[c] University of Naples Federico II Department of Electrical Engineering and Information Technologies Via Claudio 21, 80125 Naples, Italy

ARTICLE INFO

Keywords:
Cyclist safety
Active travel
Econometric models
Machine learning
Crash contributory factors
Safety countermeasures

ABSTRACT

Cycling provides numerous benefits to individuals and to society but the burden of road traffic injuries and fatalities is disproportionately sustained by cyclists. Without awareness of the contributory factors of cyclist death and injury, the capability to implement context-specific and appropriate measures is severely limited. In this paper, we investigated the effects of the characteristics related to the road, the environment, the vehicle involved, the driver, and the cyclist on severity of crashes involving cyclists analysing 72,363 crashes that occurred in Great Britain in the period 2016–2018. Both a machine learning method, as the Random Forest (RF), and an econometric model, as the Random Parameters Logit Model (RPLM), were implemented.

Three different RF algorithms were performed, namely the traditional RF, the Weighted Subspace RF, and the Random Survival Forest. The latter demonstrated superior predictive performances both in terms of F-measure and G-mean. The main result of the Random Survival Forest is the variable importance that provides a ranked list of the predictors associated with the fatal and severe cyclist crashes. For fatal classification, 19 variables showed a normalized importance higher than 5% with the second involved vehicle manoeuvring and the gender of the driver of the second vehicle having the greatest predictive ability. For serious injury classification, 13 variables showed a normalized importance higher than 5% with the bike leaving the carriageway having the greatest normalized importance. Furthermore, each path from the root node to the leaf nodes has been retraced the way back generating 361 if-then rules with fatal crash as consequent and 349 if-then rules with serious injury crash as consequent. The RPLM showed significant unobserved heterogeneity in the data finding four normal distributed indicator variables with random parameters: cyclist age $\geq$ 75 (fatal prediction), cyclist gender male (fatal and serious prediction), and driver aged 55–64 (serious prediction). The model's McFadden Pseudo R^2 is equal to 0.21, indicating a very good fit. Furthermore, to understand the magnitude of the effects and the contribution of each variable to injury severity probabilities the pseudo-elasticity was assessed, gaining valuable insights into the relative importance and influence of the variables.

The RF and the RPLM resulted complementary in identifying several roadways, environmental, vehicle, driver, and cyclist-related factors associated with higher crash severity. Based on the identified contributory factors, safety countermeasures useful to develop strategies for making bike a safer and more friendly form of transport were recommended.

* Corresponding author.
E-mail addresses: antonella.scarano@unina.it (A. Scarano), maria.rellariccardi@unina.it (M. Rella Riccardi), filomena.mauriello@unina.it (F. Mauriello), carmelo.dagostino@tft.lth.se (C. D'Agostino), nicola.pasquino@unina.it (N. Pasquino), alfonso.montella@unina.it (A. Montella).
[1] ORCIDnumber: 0000-0002-4100-453X.
[2] ORCID number: 0000-0003-2434-2577.
[3] ORCID number: 0000-0001-5682-471X.
[4] ORCID number: 0000-0003-3805-3346.
[5] ORCID number: 0000-0002-3548-299X.
[6] ORCID number: 0000-0001-9472-7056.

https://doi.org/10.1016/j.aap.2023.107275
Received 20 March 2023; Received in revised form 9 August 2023; Accepted 27 August 2023
Available online 6 September 2023

1. Introduction

In recent years, there has been a rapid rise in the bike use in most countries. In the UK, cyclist traffic grew by 62% between 2004 and 2021 (DfT, 2022a). Cycling represents a widespread and convenient form of transportation which provides numerous benefits to individuals and to society. As an active transportation mode, it is environment-friendly, associated with very little or no pollution produced (Guo et al., 2018). Furthermore, bike use is unlikely to cause death or injury to other road users (WHO, 2020). Although foregoing positive aspects, the burden of road traffic injuries and fatalities is disproportionately sustained by cyclists (WHO, 2018). It is estimated that the death and injury risks while riding a bicycle is higher than when driving a car (Nilsson et al., 2017), Thus, the cyclist is one of the road user categories with the highest crash risks.

The active transportation has attracted relevant attention because of its role in building sustainable transportation systems, in land use planning, and in profiting health. The UN Sustainable Development Goal 11, Target 11.2, promotes a shift away from motorized vehicles towards sustainable form of transport such as cycling (United Nations, 2015). It specifically requires promoting and prioritizing cycling as an accessible mean of transport for whole trips, or parts of them, and improving the infrastructure for cycling around the world. Whereby the goal can be reached, cyclist safety must be the core of global road safety and the road system must be designed and retrofitted to consider cyclists' needs.

A significant barrier to achieving uptake of cycling are the lack of cycling-friendly infrastructures as well as the road traffic danger. Improving the cyclist's safety level is a different challenge compared with motorized vehicles which have been widely exanimated in the literature (Scarano et al., 2023). Without awareness of the contributory factors of cyclist death and injury, the capability to implement context-specific and appropriate measures is limited. Thus, identifying the factors that affect crash severity is essential to improve cyclist safety.

In this paper, we investigated whether characteristics related to the road, environment, vehicle involved, driver, and cyclist were associated with cyclists' crash severity. These associations were examined using crash data from Great Britain in the period 2016–2018, implementing both a machine learning method and an econometric model. Only a few studies in the literature have used both econometric and machine learning models in a complementary way. The combination of these two approaches has allowed for the identification not only of contributing factors to the crash severity but also the most critical scenarios for cyclists.

2. Literature review

Previous studies dealing with cyclist crash severity provided evidence of the relevance of the issue for researchers, policy makers, and engineers. The researchers' commitment is trying to raise awareness on the critical factors related to roadway, environment, vehicle, crash, driver, and cyclist characteristics. Below, we tried to retrace their exertions through an in-depth literature review in the attempt to summarize the main findings about potential risk factors related to cyclist's injury severity (Table 1).

Most of the severe injuries with cyclist involvement occurred during the night and on unlit roads (Asgarzadeh et al., 2018; Boufous et al., 2012; Chang et al., 2022; Dash et al., 2022; Hosseinpour et al., 2021; Samerei et al., 2021; Wang et al., 2015) and affected the head (Brand et al., 2013; Oikawa et al., 2019). Thus, really so much research highlighted the relevance of wearing the helmet and reflective clothing to reduce fatal and serious cyclist crashes (Carlson et al., 2023; Chen and Shen, 2016; Lapparent, 2005; Marinovic et al., 2021; Moore et al., 2011; Walter et al., 2013). Rural area and higher speed limits were further connected with the increase of crash severity (Cloutier at al., 2019; Isaksson-Hellman and Töreki, 2019; Hosseinpour et al., 2021; Kaplan and Prato, 2013; Roberts and Chen, 2017; Wang et al., 2021; Yan et al.,

Table 1
Summary of the key literature findings.

Contributory factors of cyclist crash severity	References	
Roadway factors	Higher speed limits	Cloutier at al., 2019; Isaksson-Hellman et al., 2019; Kaplan and Prato, 2013; Roberts and Chen, 2017; Wang et al., 2021; Yan et al., 2011
	Rural area	Hosseinpour et al., 2021
	Unlit roads	Chang et al., 2022; Dash et al., 2022; Wang et al., 2015
	Sidewalks	Gitelman and Korchatov, 2021; Wang et al., 2021
	Curves	Alshehri et al., 2020
	Motor vehicle lanes	Du et al., 2013; Wang et al., 2021
	Lack of cycling paths separated by the motorists	Loo and Tsui, 2010; Lusk et al., 2011; Kaplan et al., 2014; Klassen et al., 2014
Environmental factors	Night-time	Asgarzadeh et al., 2018; Boufous et al., 2012; Hosseinpour et al., 2021; Samerei et al., 2021
	Foggy and rainy weather	Samerei et al., 2021; Wang et al., 2015
	Slippery and wet road surface	Kaplan et al., 2014; Rash-ha Wahi et al., 2018; Wang et al., 2015
Cyclist factors	Older cyclist	Bahrololoom et al., 2020; Behnood and Mannering, 2017; Blaizot et al., 2013; Chen and Shen., 2016; Liu et al., 2020; Oikawa et al., 2019; Samerei et al., 2021; Weber et al., 2014
	Younger cyclist (<16)	Kaplan et al., 2014; Oxley et al., 2016; Wang et al., 2015
	Male cyclist	Ouni and Belloumi, 2018; Hosseinpour et al., 2021; Meredith et al., 2020; Damsere-Derry and Bawa, 2018
	Alcohol-impaired state	Behnood and Mannering, 2017; Marinovic et al., 2021
	Not wearing a helmet or reflective clothing	Chen and Shen., 2016; Lapparent, M., 2005; Marinovic et al., 2021; Moore et al., 2011; Walter et al., 2013
	Mobile phone use when riding	Buhler et al., 2021; Du et al., 2013; Wang et al., 2021
	Red light violations	Bai and Sze, 2020; Jahangiri et al., 2016; Wu et al., 2012
	Riding on the wrong side of road	Behnood and Mannering, 2017; Hamann et al., 2015
	Speeding	Damsere-Derry and Bawa, 2018
Driver factors	Younger driver	Scholes et al.,2018
	Male driver	Mason–Jones et al., 2022; Scholes et al.,2018
	Alcohol-impaired state	Behnood and Mannering, 2017; Liu and Fan, 2021
	Hazardous overtaking	Calvi et al., 2021; Piccinini et al., 2018; Thomas et al., 2019;
	Speeding	Behnood and Mannering, 2017; Liu et al., 2020
Vehicle factors	Larger and heavier vehicle	Damsere-Derry and Bawa, 2018; Joo et al., 2017; Chen and Shen., 2016; Mason-Jones et al., 2022; Sun et al., 2022a
Crash factors	Head-on crash	Boufous et al., 2012; Lin and Fan, 2019
	Angle crash	Yan et al., 2011
	Head injury	Brand et al., 2013; Oikawa et al., 2019

2011).

Among cyclist risky behaviour, previous studies have also identified speeding (Damsere-Derry and Bawa, 2018), red light violations (Bai and Sze, 2020; Jahangiri et al., 2016; Wu et al., 2012), mobile phone use when riding (Buhler et al., 2021; Du et al., 2013; Wang et al., 2021), and riding on the wrong side of road (Behnood and Mannering, 2017; Hamann et al., 2015). On the other hand, driver hazardous overtaking and driver speeding resulted in an increase in the probability of fatal and serious crashes (Behnood and Mannering, 2017; Calvi et al., 2021; Liu et al., 2020; Piccinini et al., 2018; Thomas et al., 2019).

The psychophysical state, the age, and the gender of the cyclist further affect the cyclist' injury severity. Several studies asserted that the older and younger cyclists were associated with higher crash severity (Bahrololoom et al., 2020; Behnood and Mannering, 2017; Blaizot et al., 2013; Chen and Shen, 2016; Kaplan et al., 2014; Liu et al., 2020; Liu and Fan, 2021; Oikawa et al., 2019; Oxley et al., 2016; Samerei et al., 2021; Wang et al., 2015; Weber et al., 2014). Male cyclists were also more likely to sustain serious injuries and fatalities (Ouni and Belloumi, 2018; Hosseinpour et al., 2021; Meredith et al., 2020; Damsere-Derry and Bawa, 2018). Few studies examined the cyclist and driver psychophysical state. According with their results, alcohol-impaired state increased the crash severity (Behnood and Mannering, 2017, Liu and Fan, 2021; Marinovic et al., 2021). The features of drivers, such as gender and age also affected cyclist crash severity. Especially, previous studies found that death rates associated with male and young drivers were considerably higher (Mason-Jones et al., 2022; Scholes et al.,2018).

The severity of cyclist crashes increased when the cyclist crash occurred on the sidewalks, at curves and on the motor-vehicle lanes (Du et al., 2013; Wang et al., 2021; Gitelman and Korchatov, 2021; Alshehri et al., 2020). The availability of cycling paths separated by the motorists was found to reduce the likelihood of fatal and serious cyclist crashes (Loo and Tsui, 2010; Lusk et al., 2011; Kaplan et al., 2014; Klassen et al., 2014). As regard the environmental factors, slippery and wet road surface as well as foggy and rainy weather have been identified as contributors to the most severe cyclist crashes (Kaplan et al., 2014; Rash-ha Wahi et al., 2018; Samerei et al., 2021; Wang et al., 2015).

The size of vehicle that is involved in a cyclist crash influences the crash severity. The involvement of larger and heavier vehicles significantly increased the probability of fatal and serious consequences (Chen and Shen, 2016; Damsere-Derry and Bawa, 2018; Joo et al., 2017; Mason-Jones et al., 2022; Sun et al., 2022a). Finally, crash type was pivotal in cyclist injury severity outcome. The cyclist injury severity level could be elevated by specific crash types such as head-on and angle crashes (Boufous et al., 2012; Lin and Fan, 2019; Yan et al., 2011).

Aimed at exploring even more big data, finding structures, similarity, and concealed correlations or rules, the application of both econometric and machine learning models has widespread in recent years. The existing literature offers a variety of econometric methods to model cyclist crash severity. Methodologies include latent class cluster analysis (Akgun et al., 2018; Kent et al., 2021; Sivasankaran and Balasubramanian, 2020), logit model (Fan, 2021; Samerei et al., 2021; Salon and Mcintyre, 2018), correlated and grouped random parameters model with heterogeneity in means and/or variances (Ahmed et al., 2020; Ahmed et al., 2021; Ahmed et al., 2023; Fountas et al., 2018; Pantangi et al., 2020; Pantangi et al., 2021), negative binomial regression (Yu and Xu, 2018; Tuckel, 2021), probit model (Ghomi et al., 2016; Joo et al., 2017; Lin and Fan, 2021), and so on.

Among them, the multinomial logit model (MNL) is commonly recognized as the most extensively utilized. However, this model is burdened with several limitations that hinder its practicality, notably the IIA assumption. Thus, when the alternatives share some unobserved effects, the MNL model may not be appropriate (Lord et al., 2021; Rella Riccardi et al., 2023). To address this constraint, numerous studies have identified the Random Parameter Logit Model (RPLM) as an essentially useful tool in the analysis of discrete choices (Greene et al., 2006; McFadden and Train, 2000). It is highly flexible and allows capturing unobserved characteristics that may systematically vary across the observations. This permits to capture underlying or variable effects of independent variables that may be omitted or included in the model, respectively (Ahmed et al., 2021; Rella Riccardi et al., 2022a).

Thus, an increasing number of applications in cyclist safety field have accounted for heterogeneity into the means and variance of the distributions of the random parameters (Lin and Fan, 2021; Sun et al., 2022b; Wu et al., 2019; Ye et al., 2021b). However, the econometric models may provide unstable results due to the difficulties in handling high dimensional data. A richer set of variables can potentially improve predictive capability and understanding of causality, but it may create additional burden and complexity to the model (Mannering et al., 2020). Since the need for handling extremely large amounts of data to gain added insights in road safety while providing a high level of prediction accuracy is a current challenge that researchers are facing, we decided to combine the random parameter logit model with a data-driven tool belonging to the family of the machine learning models. Among the different machine learning approaches, the Artificial Neural Network and the Support Vector Machine exhibit high prediction performance. Nevertheless, the Artificial Neural Network and the Support Vector Machine have been criticized for their black-box nature as they suffer from a non-trivial limitation of providing outputs whose interpretability is not easy or intuitive and their results cannot be considered eligible to underly causality and select practical and safety countermeasures (Mannering et al., 2020). Even the Classification Tree (CT) has been widely used in crash severity analyses (Ghomi et al., 2016; López et al., 2014; Montella et al., 2012, 2020; Moral-Garcia et al., 2019; Prati et al., 2017), probably because it is considered the closest machine learning tool with good interpretability and ability to handle large amounts of data. Indeed, the CTs handle high-dimensional data reasonably well, ignore irrelevant descriptors making a rank of the most influential variables, and, mostly, the CTs are amenable to model interpretation due to their tree structure. The major CT drawback, however, is its relatively low prediction accuracy that may discourage its use in crash severity prediction analyses. To overcome the CTs limitation, Breiman (2001) proposed an ensembling algorithm, the Random Forest (RF), that still measures the descriptor importance, preserves the tree structure, and, mainly, provides a reduction of variance compared to the single CTs. The RF also exhibited high predictive performances (Dash et al., 2022; Komol et al., 2021; Rella Riccardi et al., 2022b; Wahab and Jiang, 2019). Such characteristics make the RF being considered particularly well suited to datasets with many features, a circumstance that is becoming more prevalent and increasingly common. In presence of large databases and a vast number of variables, the traditional classification approaches tend to become overwhelmed by the number of features and fail whereas the RF continues to perform well. To sum up, safety analysts must often consider trade-offs between the number of variables and the number of observations in the crash dataset and the intended use of the results of their research. Conscious that an ideal model would be one with excellent predictive capabilities and scalable to very large data that also uncovers causality and provides insights from crash observations, we proposed the combined use of machine learning methods and econometric models to take advantage of the different properties of the models to identify cyclists' crash severity contributory factors. In detail, in this paper we used the RF as machine learning method and the RPLM as econometric model. The RF tool was chosen among the machine learning algorithms to preserve the useful characteristics of the classification trees but with additional randomness that makes better predictions whereas the RPLM was chosen to take account of the heterogeneity among data due to unobserved characteristics that may systematically vary across the observations.

3. Crash data

The crash data used in this study have been obtained from the STATS19 dataset provided by the Department for Transport (https://

www.gov.uk). The dataset contains road crashes resulting in personal injury occurred on public highways in Great Britain. Crash data were collected by the police at the scene of the crash or were reported by a member of the public at a police station and are available at https://www.gov.uk/transport-statistics-notes-and-guidance-road-accident-and-safety. In this study, we analysed the crashes that occurred in Great Britain in the three-year period from 2016 to 2018.

In all reported crashes at least one vehicle is involved and information on Property Damage Only (PDO) crashes are not supplied. Originally, crash data were offered in three subsets storing crash, vehicle, and casualty-related information. The crash database contained thirty-two variables aimed at describing the crash event. Vehicle data included twenty-two variables describing all the vehicles involved. Finally, the casualty dataset featured sixteen variables for illustrating casualties, i.e., person injured/dead due to the crash. With the aim to work on a unique set of information, the three subsets were merged into one trough crash index, which is unique for each crash.

Furthermore, to maximize the performance of the statistical tools, the data were preliminarily prepared with appropriate transformations that consisted in the combination of some categories and the recoding of redundant information. About vehicles involved in the crash, only the bike and a second vehicle were considered for the study, since just a very small percentage of cyclist crashes involved more than two vehicles.

The final dataset consists of 72,363 cyclist crashes. It was rearranged in thirty-nine explanatory variables related to roadway (Table 2), environment (Table 3), vehicle (Table 3 and Table 4), driver (Table 5), and cyclist (Table 5). Crash severity, representing the response variable, was ranked according to the injury severity of the most seriously injured person involved in the crash and was collected into three classes: slight injury, serious injury, and fatal. It is considered fatal a crash where at least one person is killed instantly or within the thirtieth day beginning on the day in which the crash occurred. A serious injury crash is a crash resulting in an injury for which a person is held in the hospital as an "in-patient", or suffers from any injuries such as fractures, concussion, internal injuries, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment, and injuries causing death 30 or more days after the crash. Finally, a slight injury includes injuries of a minor character.

The database used in this study featured 429 fatal crashes (0.59% of the total crashes), 14,890 serious injury crashes (20.58% of the total

Table 2
Descriptive statistic related to roadway information.

Variable	Fatal		Serious		Slight		Total	
	N	%	N	%	N	%	N	%
First road class								
A[1]	213	0.29	5,865	8.10	23,298	32.20	29,376	40.60
B[2]	63	0.09	1,972	2.73	6,787	9.38	8,822	12.19
C[3]	29	0.04	930	1.29	3,889	5.37	4,848	6.70
Motorway	1	0.00	1	0.00	3	0.00	5	0.01
Missing	123	0.17	6,122	8.46	23,067	31.88	29,312	40.51
Road type								
Single carriageway	341	0.47	11,694	16.16	43,581	60.23	55,616	76.86
Dual carriageway	59	0.08	1,154	1.59	4,082	5.64	5,295	7.32
One way street	8	0.01	378	0.52	1,784	2.47	2,170	3.00
Roundabout	16	0.02	1,383	1.91	5,920	8.18	7,319	10.11
Slip road	4	0.01	83	0.11	353	0.49	440	0.61
Missing	1	0.00	198	0.27	1,324	1.83	1,523	2.10
Speed limit (mph)								
20	24	0.03	1,600	2.21	7,830	10.82	9,454	13.06
30	187	0.26	10,379	14.34	42,868	59.24	53,434	73.84
40	46	0.06	1,028	1.42	3,076	4.25	4,150	5.73
≥50	172	0.24	1,880	2.60	3,254	4.50	5,306	7.33
Missing	0	0.00	3	0.00	16	0.02	19	0.03
Junction detail								
Not at junction	240	0.33	4,858	6.71	15,212	21.02	20,310	28.07
Crossroads	144	0.20	7,,007	9.68	28,820	39.83	35,971	49.71
Other junctions	22	0.03	1,073	1.48	4,629	6.40	5,724	7.91
Roundabout	23	0.03	1,886	2.61	7,896	10.91	9,805	13.55
Missing	0	0.00	66	0.09	487	0.67	553	0.76
Junction control								
Not at junction or within 20 *m*	240	0.33	4,858	6.71	15,212	21.02	20,310	28.07
Traffic lights	57	0.08	1,479	2.04	6,374	8.81	7,910	10.93
Give way/Stop	132	0.18	8,310	11.48	33,310	46.03	41,752	57.70
Missing	0	0.00	243	0.34	2148	2.97	2,391	3.30
Second road class								
A[1]	33	0.05	1,424	1.97	6,354	8.78	7,811	10.79
B[2]	13	0.02	681	0.94	2,676	3.70	3,370	4.66
C[3]	12	0.02	581	0.80	2,741	3.79	3,334	4.61
Motorway	1	0.00	21	0.03	56	0.08	78	0.11
Missing	370	0.51	12,183	16.84	45,217	62.49	57,770	79.83
Pedestrian crossing physical facilities								
No physical crossing facilities within 50 *m*	335	0.46	11,419	15.78	40,788	56.37	52,542	72.61
Central refuge	17	0.02	458	0.63	1,896	2.62	2,371	3.28
Pedestrian phase at traffic signal junction	45	0.06	1,169	1.62	5,517	7.62	6,731	9.30
Pelican, puffin, toucan or similar non junction pedestrian light Crossing	25	0.03	1,006	1.39	3,817	5.27	4,848	6.70
Zebra	8	0.01	608	0.84	3,016	4.17	3,632	5.02
Missing	1	0.00	259	0.36	2,105	2.91	2,365	3.27

[1] *A = major roads intended to provide large-scale transport links within or between areas. Generally, an A road will be among the widest, most direct roads in an area, and is of the greatest significance to traffic travelling through the area.*

[2] *B = roads intended to connect different areas and to feed traffic between A roads and smaller roads on the network. B roads are still important routes for traffic (including traffic travelling through the area), but less so than an A road.*

[3] *C = smaller roads intended to connect together unclassified roads with A and B roads, and often linking a housing estate or a village to the rest of the network.*

Table 3
Descriptive statistics related to environment and bike information.

Variable	Fatal		Serious		Slight		Total	
	N	%	N	%	N	%	N	%
Area								
Urban	200	0.28	11,106	15.35	48,276	66.71	59,582	82.34
Rural	229	0.32	3,784	5.23	8,768	12.12	12,781	17.66
Day of week								
Weekday	306	0.42	11,489	15.88	46,367	64.08	58,162	80.38
Weekend	123	0.17	3,401	4.70	10,677	14.75	14,201	19.62
Lighting								
Daylight	317	0.44	11,647	16.10	44,798	61.91	56,762	78.44
Darkness	112	0.15	3,243	4.48	12,246	16.92	15,601	21.56
Weather								
Fine	377	0.52	13,006	17.97	48,430	66.93	61,813	85.42
Raining	33	0.05	1,158	1.60	5,022	6.94	6,213	8.59
Other	9	0.01	272	0.38	1,254	1.73	1,535	2.12
Missing	10	0.01	454	0.63	2,338	3.23	2,802	3.87
Pavement								
Dry	334	0.46	11,923	16.48	45,387	62.72	57,644	79.66
Wet/frozen	94	0.13	2,827	3.91	10,730	14.83	13,651	18.86
Missing	1	0.00	140	0.19	927	1.28	1,068	1.48
Number of bikes								
1	413	0.57	14,479	20.01	56,420	77.97	71,312	98.55
>1	16	0.02	411	0.57	624	0.86	1,051	1.45
Bike skidding and overturning								
No	357	0.49	12,599	17.41	48,250	66.68	61,206	84.58
Yes	71	0.10	1,847	2.55	4,669	6.45	6,587	9.10
Missing	1	0.00	444	0.61	4,125	5.70	4,570	6.32
Bike leaving carriageway								
No	345	0.48	13,725	18.97	51,608	71.32	65,678	90.76
Nearside	63	0.09	622	0.86	1,304	1.80	1,989	2.75
Offside	19	0.03	117	0.16	171	0.24	307	0.42
Missing	2	0.00	426	0.59	3,961	5.47	4,389	6.07
Bike hit off carriageway								
None	400	0.55	14,284	19.74	52,898	73.10	67,582	93.39
Barrier/Pole/Tree/Wall	8	0.01	120	0.17	152	0.21	280	0.39
Other	21	0.03	102	0.14	204	0.28	327	0.45
Missing	0	0.00	384	0.53	3,790	5.24	4,174	5.77
Bike 1st point of Impact								
No impact*	42	0.06	1,121	1.55	2,798	3.87	3,961	5.47
Back	105	0.15	1,457	2.01	6,449	8.91	8,011	11.07
Front	157	0.22	7,724	10.67	27,800	38.42	35,681	49.31
Nearside/Offside	124	0.17	4,263	5.89	17,303	23.91	21,690	29.97
Missing	1	0.00	325	0.45	2,694	3.72	3,020	4.17

Abbreviations: na = not admissible.
* Bike first point of impact: No impact indicates that there was no collision or contact between the bike and the vehicles, another bike, the objects, or the structures during the reported event. This mode can be useful in identifying crashes where other types of events may have occurred, such as skidding, runoff or loss of bike control, without a direct physical impact.

crashes), and 57,044 slight injury crashes (78.83% of the total crashes).

4. Method

4.1. Random forest tool

The Random Forest tool proposed by Breiman (2001) is a classifier that learns from a large number of B decision trees $\{T_1(X), \ldots, T_B(X)\}$, where $X_i = \{x_{i1}, \ldots, x_{ip}\}$ is a p-dimensional vector of the descriptors that are associated with the i_{th} crash. Each crash belongs to a dataset of N crashes, $D = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, Y_i is the crash severity class of the i_{th} crash, with $i = 1, \ldots, N$. The tree ensemble produces B outputs $\{\hat{Y}_1 = T_1(X), \ldots, \hat{Y}_B = T_B(X)\}$, where $\hat{Y}_b$, $b = 1, \ldots, B$, is the prediction for the i_{th} crash generated by the b_{th} tree. After a large number of trees is generated, the outputs are aggregated to produce one final prediction $\hat{Y}$ which is the most popular class voted by the majority of the trees.

Overtime, multiple algorithms have been developed to carry out RF analyses with the aim of improving the performances of the classifier especially in presence of many features (as the GB database that contains 39 variables and 72,363 crashes with cyclist involvement) and with a small proportion of the true cases (fatal and serious cyclist crashes in this study). In such circumstances, the traditional RF tool performance tends to decline with respect to the target classification (Amaratunga et al., 2008) as informative variables have too little opportunity of being selected among all the available variables. In this study, we performed three different RF algorithms: the traditional RF (Breiman, 2001), the Weighted Subspace RF (Xu et al., 2012), and the Random Survival Forest (Ishwaran et al., 2008). In the traditional RF, each tree of the forest is trained only on a sample of N crashes drawn at random with replacement from the complete dataset of N crashes. This procedure is also known as the bagging step and the selected samples are called the in-bag cases. Instead of simple random sampling, the Weighted Subspace RF uses a weighting method for feature subspace selection to enhance classification performance over high-dimensional data. Finally, the Random Survival Forest algorithm uses and "ensemble learning" that be resumed in the following steps:

1) it draws a bootstrap sample with random feature selection and with a replacement from the original sample;
2) for each bootstrap sample, it grows a tree and chooses the best split among a randomly selected subset of descriptors at each node; at each node, the variable selected as the best splitter groups the observations into disjoint classes which are externally heterogeneous and internally homogeneous;

Table 4
Descriptive statistics related to vehicle information.

Variable	Fatal		Serious		Slight		Total	
	N	%	N	%	N	%	N	%
Vehicle 2 skidding and overturning								
No	336	0.46	12,638	17.46	49,920	68.99	62,894	86.91
Yes	21	0.03	241	0.33	520	0.72	782	1.08
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	1	0.00	380	0.53	3,754	5.19	4,135	5.71
Vehicle 2 leaving carriageway								
No	330	0.46	12,665	17.50	50,066	69.19	63,061	87.15
Nearside	12	0.02	150	0.21	388	0.54	550	0.76
Offside	15	0.02	80	0.11	152	0.21	247	0.34
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	1	0.00	364	0.50	3,588	4.96	3,953	5.46
Vehicle 2 hit off carriageway								
None	346	0.48	12,849	17.76	50,652	70.00	63,847	88.23
Barrier/Pole/Tree/Wall	9	0.01	54	0.07	83	0.11	146	0.20
Other	3	0.00	35	0.05	59	0.08	97	0.13
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	0	0.00	321	0.44	3,400	4.70	3,721	5.14
Vehicle 2 1st point of impact								
No impact	10	0.01	860	1.19	2,680	3.70	3,550	4.91
Back	13	0.02	807	1.12	3,202	4.42	4,022	5.56
Front	246	0.34	6,251	8.64	25,419	35.13	31,916	44.11
Nearside/Offside	89	0.12	5,087	7.03	20,569	28.42	25,745	35.58
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	0	0.00	254	0.35	2,324	3.21	2,578	3.56
Vehicle 2 engine capacity								
≤1000	26	0.04	749	1.04	3024	4.18	3799	5.25
1001–1500	54	0.07	2,809	3.88	11,413	15.77	14,276	19.73
1501–2000	106	0.15	4,428	6.12	18,572	25.67	23,106	31.93
2001–3000	38	0.05	1,359	1.88	5,301	7.33	6,698	9.26
>3000	78	0.11	520	0.72	1439	1.99	2,037	2.81
Missing	50	0.07	3,165	4.37	14,140	19.54	17,355	23.98
na	77	0.11	1,860	2.57	3,155	4.36	5,092	7.04
Vehicle 2 propulsion code								
Petrol	113	0.16	4,979	6.88	20,294	28.04	25,386	35.08
Heavy oil	189	0.26	4,626	6.39	18,015	24.90	22,830	31.55
Other	3	0.00	305	0.42	1,583	2.19	1,891	2.61
na	77	0.11	1,860	2.57	3,155	4.36	5,092	7.04
Missing	47	0.06	3,120	4.31	13,997	19.34	17,164	23.72
Vehicle 2 age								
≤15	281	0.39	9,132	12.62	36,775	50.82	46,188	63.83
>15	19	0.03	652	0.90	2,412	3.33	3,083	4.26
Missing	47	0.06	3,077	4.25	13,725	18.97	16,849	23.28
na	82	0.11	2,029	2.80	4,132	5.71	6,243	8.63
Vehicle 2 type								
Car	213	0.29	10,839	14.98	46,538	64.31	57,590	79.58
Bike	6	0.01	229	0.32	305	0.42	540	0.75
Bus	17	0.02	275	0.38	975	1.35	1,267	1.75
PTW	16	0.02	283	0.39	1,041	1.44	1,340	1.85
Truck	66	0.09	381	0.53	972	1.34	1,419	1.96
Van	29	0.04	1,064	1.47	3,808	5.26	4,901	6.77
Other	11	0.02	129	0.18	400	0.55	540	0.75
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	0	0.00	59	0.08	155	0.21	214	0.30
Vehicle 2 towing and articulation								
No	322	0.44	12,674	17.51	52,343	72.33	65,339	90.29
Articulated vehicle	20	0.03	90	0.12	121	0.17	231	0.32
Other	9	0.01	115	0.16	310	0.43	434	0.60
na	77	0.11	1,860	2.57	3,155	4.36	5,092	7.04
Missing	1	0.00	151	0.21	1,115	1.54	1,267	1.75
Vehicle 2 manoeuvre								
Going ahead	260	0.36	5,099	7.05	17,646	24.39	23,005	31.79
Moving off	9	0.01	1,078	1.49	5,254	7.26	6,341	8.76
Overtaking	30	0.04	764	1.06	2,857	3.95	3,651	5.05
Turning left/right/U/Reversing	37	0.05	4,235	5.85	18,099	25.01	22,371	30.91
Other	21	0.03	1,710	2.36	6,670	9.22	8,401	11.61
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	1	0.00	373	0.52	3,668	5.07	4,042	5.59

Abbreviations: na = not admissible, PTW = Powered two-wheeler.

Table 5
Descriptive statistics related to cyclist and driver information.

Variable	Fatal		Serious		Slight		Total	
	N	%	N	%	N	%	N	%
Cyclist manoeuvre								
Going ahead	362	0.50	11,753	16.24	42,128	58.22	54,243	74.96
Moving off	9	0.01	360	0.50	1,696	2.34	2,065	2.85
Overtaking	7	0.01	633	0.87	2,462	3.40	3,102	4.29
Turning left/right/U/Reversing	37	0.05	1100	1.52	4,011	5.54	5,148	7.11
Other	13	0.02	635	0.88	2,947	4.07	3,595	4.97
Missing	1	0.00	409	0.57	3,800	5.25	4,210	5.82
Cyclist journey purpose								
Commuting to/from work	55	0.08	2,724	3.76	11,340	15.67	14,119	19.51
Journey as part of work	12	0.02	793	1.10	3,351	4.63	4,156	5.74
To/from school	6	0.01	327	0.45	2,047	2.83	2,380	3.29
Other	45	0.06	1,647	2.28	4,204	5.81	5,896	8.15
Missing	311	0.43	9,399	12.99	36,102	49.89	45,812	63.31
Cyclist gender								
Female	58	0.08	2,559	3.54	11,120	15.37	13,737	18.98
Male	369	0.51	12,201	16.86	45,433	62.78	58,003	80.16
Missing	2	0.00	130	0.18	491	0.68	623	0.86
Cyclist age								
≤17	36	0.05	1,710	2.36	8,331	11.51	10,077	13.93
18–24	28	0.04	1,432	1.98	6,825	9.43	8,285	11.45
25–34	52	0.07	2,651	3.66	13,004	17.97	15,707	21.71
35–44	51	0.07	2,616	3.62	10,050	13.89	12,717	17.57
45–54	72	0.10	2,897	4.00	8,788	12.14	11,757	16.25
55–64	74	0.10	1,679	2.32	4,228	5.84	5,981	8.27
65–74	55	0.08	675	0.93	1,379	1.91	2,109	2.91
≥75	36	0.05	278	0.38	462	0.64	776	1.07
Missing	25	0.03	952	1.32	3,977	5.50	4,954	6.85
Cyclist IMD								
Less deprived	198	0.27	6,291	8.69	21,527	29.75	28,016	38.72
More deprived	154	0.21	6,637	9.17	28,020	38.72	34,811	48.11
Missing	77	0.11	1,962	2.71	7497	10.36	9,536	13.18
Cyclist home area								
Urban	247	0.34	11,039	15.26	44,616	61.66	55,902	77.25
Rural	69	0.10	1,065	1.47	2,585	3.57	3,719	5.14
Small town	36	0.05	825	1.14	2,348	3.24	3,209	4.43
Missing	77	0.11	1,961	2.71	7,495	10.36	9,533	13.17
Driver 2 journey purpose								
Commuting to-from work/school	27	0.04	1,653	2.28	5,719	7.90	7,399	10.22
Journey as part of work	122	0.17	2,193	3.03	7,676	10.61	9,991	13.81
Other	37	0.05	1,254	1.73	3,417	4.72	4,708	6.51
na	71	0.10	1,631	2.25	2850	3.94	4552	6.29
Missing	172	0.24	8,159	11.28	37,382	51.66	45,713	63.17
Driver 2 gender								
Female	55	0.08	3,452	4.77	14,031	19.39	17,538	24.24
Male	287	0.40	7,766	10.73	29,765	41.13	37,818	52.26
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	16	0.02	2,041	2.82	10,398	14.37	12,455	17.21
Driver 2 age								
≤17	3	0.00	94	0.13	315	0.44	412	0.57
18–24	45	0.06	1,160	1.60	3,902	5.39	5,107	7.06
25–34	75	0.10	2,330	3.22	8,852	12.23	11,257	15.56
35–44	66	0.09	2,004	2.77	8,077	11.16	10,147	14.02
45–54	61	0.08	2,118	2.93	8,075	11.16	10,254	14.17
55–64	57	0.08	1,489	2.06	5,279	7.30	6,825	9.43
65–74	16	0.02	779	1.08	2,837	3.92	3,632	5.02
≥75	13	0.02	616	0.85	1,826	2.52	2,455	3.39
na	76	0.11	1,817	2.51	3,850	5.32	5,743	7.94
Missing	17	0.02	2,483	3.43	14,031	19.39	16,531	22.84
Driver 2 IMD								
Less deprived	135	0.19	4,667	6.45	16,488	22.79	21,290	29.42
More deprived	160	0.22	4,926	6.81	19,578	27.06	24,664	34.08
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	63	0.09	3,666	5.07	18,128	25.05	21,857	30.20
Driver 2 home area								
Urban	203	0.28	7,785	10.76	30,728	42.46	38,716	53.50
Rural	57	0.08	1,013	1.40	3,007	4.16	4,077	5.63
Small town	35	0.05	796	1.10	2,333	3.22	3,164	4.37
na	71	0.10	1,631	2.25	2,850	3.94	4,552	6.29
Missing	63	0.09	3,665	5.06	18,126	25.05	21,854	30.20
Crash severity	**429**	**0.59**	**14,890**	**20.58**	**57,044**	**78.83**	**72,363**	**100.00**

Abbreviations: IMD = Index of Multiple Deprivation; na = not admissible.

3) it repeats the above steps until B trees are generated.

In this study, we used the Classification And Regression Tree (CART) (Breiman et al., 1984) algorithm to build the forest. At each node, the CART algorithm selects the variable candidate for splitting according to the maximum decrease in the impurity of the node. The impurity (or heterogeneity) of each node is assessed through the Gini reduction criterion (the higher the value of the Gini index, the higher the homogeneity of the node that is due to the split) which can be calculated as follows (López et al., 2014; Montella et al., 2011):

$$i_Y(t) = 1 - \sum_j p(j|t)^2 \quad (1)$$

where:

$i_Y(t)$ is the impurity of the node t,

$p(j|t)$ is the proportion of crashes in the node t that belong to the class j.

It has been shown that there is a potential overestimate of the true prediction error, depending on the choices of the random forest hyperparameters, such as the number of trees (B), and the number of descriptors. To reduce the true prediction error, the out-of-bag estimate of the error rate (ER_{OOB}) was estimated by varying the B and the number of descriptors:

$$ER^{OOB} = \frac{\sum_{i=1}^{N} \left(\widehat{Y}^{OOB}(X_i) \neq Y_i\right)}{N} \quad (2)$$

where:

$\widehat{Y}^{OOB}(X_i)$ is the predicted class for the i_{th} crash;

X_i is the vector of the descriptors that are associated with the i_{th} crash;

Y_i is the crash severity class of the i_{th} crash;

N is the total number of crashes in the database.

The values of the number of trees and the number of descriptors were chosen so that the ER^{OOB} tends to stabilize around the minimum value.

For each tree in the "forest", a set of randomly selected variables governs the tree growth. The importance of each variable determines the predictive power of the variable itself in the forest. The contribution of the variable x_j to the homogeneity of each node in a particular tree is:

$$VI = \sum_{t=1}^{T} \frac{N(t)}{N} \Delta i_Y(t,s) \quad (3)$$

where:

VI is the relative importance of variable x_j;

$\Delta i_Y(t,s)$ is the reduction in the Gini index obtained by splitting variable x_j at node t;

N is the total number of observations;

T is the number of nodes in the tree.

The variable importance evaluated for the variable x_j, $(VI(x_j))$, is computed as the sum of the importance over all of the B trees in the forest:

$$VI(x_j) = \frac{\sum_{b=1}^{B} VI^b(x_j)}{B} \quad (4)$$

where:

$VI^b(x_j)$ is the variable importance of the b_{th} tree, calculated using the equation (2);

B is the number of trees.

The RF was performed in the R-CRAN software environment using the packages, "randomForest", "wsrf", and "randomForestSRC".

To further explore the random forest tool and understand how the patterns are combined to make predictions, each path from the root node to the leaf nodes of each generated tree of the forest has been retraced the way back to generate if-then rules where the splitters from the root node to the terminal node are the antecedents and the severity class of the terminal node is the consequent. The procedure is inspired to the association rule discovery technique (Das et al., 2019; Montella et al., 2021), another machine learning tool that highlights items that occur frequently together in a crash dataset. Each crash record contains different items (e.g., road type, area, day of week, crash severity, pavement, …) and the dataset contains all the items of each crash. Basing on the relative frequency of the number of times the sets of items occur alone and in combination in a dataset, the association rules were extracted with the form "A → B", where A and B are disjoint item-sets: A is the antecedent and B is the consequent. The conversion of the trees in the forest in if-then rules allows to visualize the different patterns that are mostly combined each other when a cyclist crash occur and aims to identify potential scenarios that are critical in the interaction between bikes and the other motorized vehicles.

Gaining results that identify the factors contributing the most to the cyclist crashes while accurately predicting the true cases (fatal and serious cyclist crashes in this study) is highly desirable. Hence, we carried out three different RF tools and then their results were compared in terms of prediction accuracy measured with F-measure and G-mean indicators (paragraph 4.3).

4.2. Random parameters logit model

The random parameter logit model (also known as mixed logit model) belongs to the econometric model. It is the generalized version of the standard multinomial logit model and allows the estimated parameters β_j to be fixed or to be random across all the observations in the database of size N. Such randomness can be expressed in the evaluation of injury-severity functions S_{ij}:

$$S_{ij} = \beta_j X_{ij} + \varepsilon_{ij} \quad (5)$$

where:

S_{ij} represents the propensity for a crash i_{th} of being recorded with the crash severity class j (here j varies from 1 to 3: 1 = slight injury, 2 = serious injury, 3 = fatal);

X_{ij} is the vector of explanatory variables that are associated with the i_{th} crash;

β_j is the column vector of the estimable parameters for the crash severity class j;

ε_{ii} is the error term, independent and identically distributed, following the Type I generalized extreme value distribution (i.e., Gumbel), and independent of underlying parameters or data characteristics;

σ is the vector of parameters characterizing the distribution of β_j in terms of mean and variance.

The random parameter multinomial logit model overcomes the multinomial logit model limitations by avoiding issues regarding IIA (Independence of Irrelevant Alternatives) violation and permitting the heterogeneous effects and correlation in unobserved factors. Thus, if unobserved heterogeneity is allowed, β_j is a vector with a continuous density function $f(\beta|\sigma)$ and the injury severity outcome probabilities are defined as follows:

$$P_{ij} = \int \frac{e^{\beta_j X_{ij}}}{\sum_J e^{\beta_j X_{ij}}} f(\beta|\sigma) d\beta \quad (6)$$

The injury severity outcome probability is purely a weighted average for different values of β_j across cyclist crashes where the elements of the vector β_j may be fixed or randomly distributed (Anastasopoulos and Mannering, 2011).

To estimate how the model fits the data, the McFadden's Pseudo R^2 was assessed:

$$McFaddenR^2 = 1 - \frac{LL_{full}}{LL_0} \quad (7)$$

where:

LL_0 is the loglikelihood of the null model ;

LL_{full} is the loglikelihood of the model including all statistically significant variables.

The McFadden's Pseudo R^2 variability range is between 0 and 1. However, McFadden's Pseudo R^2 greater than 0.20 indicates a very good fit (Andreß et al., 2013).

Before implementing the model, all variables were tested for multicollinearity by using the Pearson's correlation coefficient (Pearson, 1896). It measures the degree of linear relationship between two variables. A correlation coefficient of + 1 or − 1 indicates a perfect positive or negative linear correlation, respectively. Variables are considered strongly correlated if the absolute value of Pearson's correlation is higher than 0.85 and the estimated correlation has p-value less than 0.05 (Cafiso et al., 2010; Montella and Imbriani, 2015). Variables included in the models are not highly correlated.

The RPLM was performed in the R-CRAN software environment using the package "mlogit".

The regression coefficients derived from RPLM merely provide a qualitatively impact of significant independent variables on crash severity. This implies that the model coefficients cannot be directly interpreted. Thus, to paraphrase the quantitatively impact of specific variables on the injury-outcome probabilities the elasticity should be computed. In the case of a continuous variable, elasticity represents the percentage change in the outcome when the predictor variable increases by 1% (Lord et al., 2021). It is obtained by calculating the partial derivative with respect to the continuous variable for each individual observation. However, in our study explanatory variables were categorical, thus they were transformed into dummy variables. Each unique value of the original categories was assigned a corresponding dummy variable. When the independent variable is a dummy variable, pseudo-elasticity must be employed to evaluate the impact of individual parameter estimates on the probabilities of crash severity. The specific pseudo-elasticities can be calculated using the following equation (Ye et al., 2021a):

$$E_{X_{im}}^{P_{ji}} = \left[\exp(\beta_{jm}) \frac{\sum_J \exp(ASC_j + \beta_j X_i)}{\sum_J \exp(\Delta(ASC_j + \beta_j X_i)} - 1\right] \times 100 \quad (8)$$

where:

$E_{X_{im}}^{P_{ji}}$ is the direct pseudo-elasticity of the m_{th} variable from the vector X_i;

J is the number of crash severity class;

$ASC_j + \beta_j X_i$ is the function value that determines the crash severity class when X_{im} is zero;

$\Delta(ASC_j + \beta_j X_i)$ is the function value after modifying X_{im} from zero to one.

The pseudo-elasticity of an indicator variable, in relation to a crash severity category, reflects the percentage alteration in the probability of that specific crash severity class when the variable is shifted from zero to one.

4.3. Performance metrics

The performance of the RF algorithms and the RPLM were evaluated considering the G-mean and the F-measure, which are multi-parameter indicators that combine more simple indicators into a single performance measure, and the False Positive rate. The F-measure value combines the precision and the recall, both indicators of correct classification of the true cases. A high value of F-measure indicates high prediction accuracy in classifying the fatal and serious injury crashes. G-mean combines the performance of the model in correctly classifying both the positive and negative classes. The performance measures were evaluated as follows (Guo et al., 2008):

$$\mathrm{F-measure} = \frac{(1+\beta^2) \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (9)$$

where:

$\mathrm{Precision} = \frac{TP}{TP+FP}$ is the probability of correct classification;

$\mathrm{Recall} = \mathrm{Acc}^+ = \frac{TP}{TP+FN}$ is the true positive rate. TP is the number of true positives; FN is the number of false negatives.

β is a coefficient to adjust the relative importance of precision versus recall.

β was set equal to 1 to equally consider precision and recall (Bekkar et al., 2013).

$$\mathrm{G-mean} = (\mathrm{Acc}^- \times \mathrm{Acc}^+)^{\frac{1}{2}} \quad (10)$$

where:

$\mathrm{Acc}^- = \frac{TN}{TN+FP}$ is the true negative rate; TN is the number of true negatives; FP is the number of false positives.

The False Positive rate (FP_{rate}) indicates how often the model is likely to miss-predict. As shown by previous studies, it is an important metric for unbalanced datasets (Islam and Abdel-Aty, 2023; Li et al., 2020):

$$FP_{rate} = \frac{FP}{FP + TN} \quad (11)$$

5. Results

5.1. Random forest tool

Due to the considerable number of features collected in the GB database and the true cases being only a small proportion of the total crashes, the performances exhibited by the traditional RF tool and the Weighted Subspace RF were poor, mainly when it comes to classify the fatal and serious crashes (Table 6). The traditional RF and the Weighted Subspace RF correctly classify few fatal and serious injury crashes. Also, the traditional RF and the Weighted Subspace RF required a number of trees sufficiently large to obtain stable results (1500 trees and 500 trees to make up the forest respectively) with a computational burden higher than the computational burden required to perform the Random Survival Forest algorithm. The Random Survival Forest was initially carried out generating 500 trees. Then, the hyperparameter tuning process provided the optimal number of trees equal to 68. The error rate was significantly smaller than 0.5, the benchmark value associated with a procedure no better than flipping a coin (Ishwaran and Kogalur, 2007). The tree depth was set equal to 4 levels.

The Random Survival Forest algorithm is the most straightforward RF algorithm able to capture the complex relationships between the predictors and cyclist crash severity showing superior predictive performances both in terms of F-measure, G-mean, and FP_{rate}. Hence, only the Random Survival Forest results were provided.

The main result of the Random Survival Forest is the variable importance that provided a ranked list of the predictors associated with the fatal and severe cyclist crashes. The most important contributors exhibited high importance whereas variables whose importance is close

Table 6
Performance measures of the RF models.

	Number of trees	F-measure	G-mean	FP_{rate}
Traditional RF	**1500**			
Fatal		0.00	0.00	0.00
Serious injury		0.09	0.22	0.01
Slight injury		0.88	0.22	0.95
Weighted Subspace RF	**500**			
Fatal		0.01	0.06	0.00
Serious injury		0.16	0.31	0.04
Slight injury		0.87	0.30	0.90
Random Survival Forest	**Firstly 500, then 68**			
Fatal		0.09	0.74	0.07
Serious injury		0.28	0.48	0.21
Slight injury		0.78	0.51	0.56

to zero are variables that contribute nothing to prediction. There are some variables with negative values of importance, whose presence is associated to a noise in the classification process.

Unfortunately, the tool cannot directly correlate crash contributory factors with crash severity, nor it can be used to quantify the impact of the contributing factors on injury severity. As an attempt to overcome these drawbacks of the RF tool, the if-then rules that had been generated for each tree in the forest by retracing the way back each path from the root node to the leaf nodes and were below provided. The splitters (the predictors giving the best partition at each level of the tree) are the antecedents whereas the severity class of the terminal node is the consequent. The RF provided 1,038 rules. The generated rules were ordered by the decreasing proportion of crashes that each rule covered in the related tree. Then, the most relevant rules were selected according to the highest frequency of appearance in the trees. The 20 more frequent rules having fatal cyclist crashes as consequent were reported in Table 7, whereas the 20 more frequent rules having serious injury cyclist crashes as consequent were reported in Table 8. The conversion of the trees of the forest in if-then rules allows to visualize how the different patterns are likely to be combined each other when a cyclist crash occur and aims to identify potential scenarios that are critical in the interaction between the bikes and the other motorized vehicles.

5.1.1. Fatal variable importance and if-then rules

Fig. 1 shows the normalized importance of the variables in the forest for fatal classification. The variables with normalized importance higher than 5% were 19, with the second involved vehicle manoeuvring (vehicle 2 manoeuvre) and the gender of the driver of the second vehicle (driver 2 gender) having the greatest normalized importance (at least 50% of importance in the classification process) and higher predictive ability. The variable vehicle 2 manoeuvre included going ahead, moving off, overtaking, turning left, right, U inversion and reversing while the variable driver 2 gender provided information about the greater propensity of being involved in a cyclist crash based on the gender of the driver of the second involved vehicle in case of multi-vehicle crashes. The moderately important variables (importance in the classification process around 20–50%) were 5 and included: (1) area, that contributed to understand if the fatal crash with cyclist involvement occurred in urban or rural area; (2) junction control, that provided information about the fatal crash related to the intersection organization (traffic lights or give way/stop) if the crash occurred at intersection; (3) cyclist manoeuvre, that included the same manoeuvres of the second vehicle involved; (4) the journey purpose of the driver of the second vehicle involved, that provided information related to the driver purpose of travel, if commuting to-from work/school or the journey as part of work; and (5) junction detail, that contributed to understand if the crash occurred at crossroads or roundabouts in case of crashes at intersections. There were also 12 variables of small importance (less than 20% of importance).

The fatal crash was the consequent for 361 rules (1 with two antecedents, 15 with three antecedents, 345 with four antecedents, 34.8% of the total rules). As identified by the variable importance, the manoeuvre of the second vehicle involved in the cyclist crash is pivotal. In the if-then rules, the variable occurred as the first item in 57 rules.

Moreover, the rules also identified the type of the second vehicle involved as well as its engine capacity having great influence on the severity of the cyclist crashes. The type of the second vehicle involved generated 55 rules as first antecedent whereas the vehicle engine capacity 49. In many cases, the RF tool found a combination of patterns that occurred together in fatal cyclist crashes. For instance, the rule T47_2 associated the combination of three antecedents, namely (1) the second vehicle manoeuvre equal to going ahead or overtaking, (2) the second vehicle type equal to car or van, and (3) the area equal to rural, with the fatal severity. To better understand the meaning of this rule, consider that the manoeuvre of the second vehicle involved (going ahead/overtaking), the second vehicle type (car/van), and the area

Table 7
Top 20 if-then rules for fatal cyclist crashes generated by RF.

ID Rule	Antecedents	Consequent	#	%
T3_5	Vehicle 2 leaving carriageway = No, Nearside & Driver 2 age ≤ 17, 18–24, 25–34, 35–44, 45–54, ≥75 & Junction control = Traffic lights, No junction	Fatal	273	29.84
T34_14	Vehicle 2 manoeuvre = Going ahead, Overtaking, Turning left/right/U/ Reversing & Vehicle 2 type = Bike, PTW, Car, Bus, & Driver 2 age = 18–24, 25–34, 55–64	Fatal	222	24.75
T19_12	Driver 2 age = 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, ≥75 & Junction detail = Roundabout & Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Vehicle 2 type = PTW, Car, Van, Bus	Fatal	215	23.97
T26_14	Vehicle 2 manoeuvre = Going ahead, Overtaking & Driver 2 age ≤ 17, 18–24, 25–34, 35–44, 45–54, 55–64 & Vehicle 2 engine capacity ≤ 1000, 1001–1500, 2001–3000, >3000 & Cyclist home area = Urban	Fatal	214	23.86
T57_11	Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Area = Rural & Road type = Dual carriageway, One-way street, Single carriageway, Slip road & Vehicle 2 skidding and overturning = No	Fatal	172	19.39
T63_8	Vehicle 2 manoeuvre = Going ahead & Cyclist age = 25–34, 35–44, 65–74	Fatal	166	18.57
T63_5	Vehicle 2 manoeuvre = Going ahead & Cyclist age = 18–24, 45–54, 55–64, ≥75	Fatal	158	17.67
T65_5	Vehicle 2 towing and articulation = No & Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Junction control = Traffic lights, No junction & Driver 2 age = 18–24, 25–34, 45–54, 55–64, 65–74	Fatal	154	17.17
T60_7	Area = Rural & Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Bike hit off carriageway = Barrier/ Pole/Tree/Wall & Driver 2 age = 25–34, 45–54, 55–64, 65–74, ≥75	Fatal	148	16.50
T18_10	Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Vehicle 2 type = PTW, Car, Van & Area = Rural	Fatal	144	16.05
T23_14	Bike 1st Point of impact = Back, Nearside/Offside & Vehicle 2 manoeuvre = Going ahead, Moving off, Overtaking & Junction detail = No junction & Driver 2 age = 18–24, 25–34, 35–44, 45–54, 55–64, 65–74	Fatal	144	16.05
T24_11	Vehicle 2 hit off carriageway = Barrier/ Pole/Tree/Wall & Driver 2 Gender = M & Cyclist home area = Rural, Small town & Driver 2 age ≤ 17, 18–24, 25–34, 35–44, 45–54, 55–64, ≥75	Fatal	139	15.50
T40_8	Area = Rural & Vehicle 2 manoeuvre = Going ahead & Junction detail = Crossroad, Roundabout, No junction	Fatal	131	14.60
T62_7	Area = Rural & Vehicle 2 manoeuvre = Going ahead, Turning left/right/U/ Reversing & Vehicle 2 type = Bike, Car, Van, Bus & Bike leaving carriageway = No	Fatal	129	14.38
T30_7	Bike leaving carriageway = No & Driver 2 Gender = M & Area = Rural	Fatal	127	14.16
T39_10	Vehicle 2 type = Car, Van, Bus & Bike leaving carriageway = No & Cyclist home area = Rural, Small town & Vehicle 2 manoeuvre = Going ahead, Overtaking	Fatal	118	14.10

(continued on next page)

Table 7 (*continued*)

ID Rule	Antecedents	Consequent	#	%
T27_6	Bike leaving carriageway = No & Cyclist home area = Rural, Small town & Vehicle 2 manoeuvre = Going ahead, Overtaking & Vehicle 2 type = Bike, PTW, Car, Van, Truck	Fatal	124	13.82
T59_14	Cyclist home area = Rural, Small town & Vehicle 2 towing and articulation = No & Area = Rural	Fatal	115	12.82
T32_11	Vehicle 2 engine capacity ≤ 1000, 1001–1500, 1501–2000, >3000 & Driver 2 Journey Purpose = Commuting to-from work/school, Other & Vehicle 2 manoeuvre = Going ahead, Moving off & Cyclist age = 18–24, 45–54, 55–64, 65–74, ≥75	Fatal	112	12.49
T47_2	Vehicle 2 manoeuvre = Going ahead, Overtaking & Vehicle 2 type = Car, Van & Area = Rural	Fatal	109	12.21

(rural) are the splitter variables. Their combined presence increases the fatal severity propensity in cyclist crashes. Hence, this association of factors shall be considered a potential cause of cyclist crashes resulting in fatal outcome.

In 51 rules, the most frequent manoeuvres associated to the second vehicle involved in the cyclist crash were the turning left, turning right, U inversion, and reversing. The vehicle going ahead, proceeding on its route occurred in 40 rules. As the vehicle type, bus and truck occurred in 56 rules, followed by the PTW which was found in 39 rules. The rule T34_14 identified that the patterns of the second vehicle involved in a cyclist crash are associated with a greater propensity to fatal severity. The driver gender of the second vehicle involved in the crash only generated 7 rules as the first antecedent but it was often associated with other variables in 27 rules. Of which, 17 were with male drivers' involvement. Area was found in 41 rules as first antecedent. Of which, 25 reported the rural context as the most frequent antecedent in fatal cyclist crashes.

The strongest rule identified by the RF for fatal crashes (T3_5) covered roughly 30% of the observations falling in the tree from which the rule was extracted. Overall, the rule covered roughly 64% of the total fatal crashes in the database and contained the second vehicle involved in the crashes that does not leave the carriageway or those leaving the nearside of the carriageway. The age of the drivers of the second vehicles are young drivers (up to 44 years old), middle-aged (between 45 and 54 years old), and very elderly drivers (at least 75 years old) that fail to navigate a traffic light intersection hitting a cyclist or hit the cyclist far from the intersections. An interesting rule is the rule T57_11 covering those fatal cyclist crashes that occurred in rural area because of the second vehicle involved that goes ahead, moves off or overtakes another vehicle on road segments, far from intersections. The rule T23_14 associated the manoeuvres of the second vehicle (going ahead, moving off, and overtaking) on road segments, far from intersections, and the second vehicle involved that hits the cyclist on the back or nearside/offside with the fatal crash as a consequence. Another interesting rule (T32_11) identified the association between commuting to-from work or school as the purpose of the journey of the driver of the second vehicle involved and the fatal cyclist crash.

5.1.2. Serious injury variable importance and if-then rules

Fig. 2 shows the normalized importance of the variables in the forest for serious injury classification. The variables with normalized importance higher than 5% were 13, with the bike leaving the carriageway having the greatest normalized importance. If the cyclist leaved the carriageway because of the crash, the variable provides understanding on the direction of leaving, nearside or offside. The moderately important variables (importance in the classification process around 20–50%)

Table 8
Top 20 if-then rules for serious injury crashes generated by RF.

ID Rule	Antecedents	Consequent	#	%
T54_12	Bike leaving carriageway = No & Vehicle 2 engine capacity ≤ 1000, 1001–1500, 1501–2000, 2001–3000 & Vehicle 2 towing and articulation = No & Driver 2 age = 18–24, 25–34, 35–44, 55–64, 65–74, ≥75	Serious	449	50.06
T39_8	Vehicle 2 type = PTW, Car, Van, Bus & Bike leaving carriageway = No & Cyclist home area = Urban	Serious	406	48.51
T27_2	Bike leaving carriageway = No & Cyclist home area = Urban & Vehicle 2 skidding and overturning = No	Serious	430	47.94
T30_8	Bike leaving carriageway = No & Driver 2 Gender = M & Area = Urban	Serious	401	44.70
T24_9	Driver 2 Gender = M & Cyclist home area = Urban & Vehicle 2 towing and articulation = No	Serious	341	38.02
T3_6	Vehicle 2 leaving carriageway = No, Nearside & Driver 2 age ≤ 17, 18–24, 25–34, 35–44, 45–54, ≥75 & Junction control = Give way/Stop	Serious	323	35.30
T67_14	Vehicle 2 type = Bike, PTW, Car, Van & Vehicle 2 manoeuvre = Going ahead, Overtaking	Serious	311	34.67
T48_12	Area = Urban & Vehicle 2 type = Bike, PTW, Car, Van, Bus & Driver 2 age ≤ 17, 18–24, 35 44, 45–54, 55–64, ≥75	Serious	307	34.23
T23_7	Bike 1st point of impact = Front & Area = Urban	Serious	286	31.88
T49_9	Vehicle 2 manoeuvre = Going ahead & Vehicle 2 type = Car, Van, Bus	Serious	265	29.54
T13_12	Driver 2 Gender = M & Cyclist home area = Urban & Vehicle 2 engine capacity ≤ 1000, 1001–1500, 1501–2000 & Cyclist manoeuvre = Going ahead, Overtaking, Turning left/right/U/Reversing	Serious	256	28.54
T4_14	Bike leaving carriageway = No & Vehicle 2 type = Bike, PTW, Car, Van & Junction detail = No junction, Other & Vehicle 2 manoeuvre = Going ahead, Overtaking, Turning left/right/U/Reversing	Serious	242	26.45
T66_16	Vehicle 2 engine capacity ≤ 1000, 2001–3000, >3000 & Vehicle 2 type = Bike, PTW, Car, Van & Speed limit = 30, 40	Serious	226	25.20
T43_15	Vehicle 2 type = Bike, PTW, Car, Van & Driver 2 age ≤ 17, 18–24, 25–34, 35–44, 55–64, 65–74, ≥75 & Vehicle 2 manoeuvre = Going ahead	Serious	206	24.91
T34_15	Vehicle 2 manoeuvre = Going ahead, Overtaking, Turning left/right/U/Reversing & Vehicle 2 type = Bike, PTW, Car, Bus & Driver 2 age ≤ 17, 35–44, 45–54, 65–74, ≥75	Serious	210	23.41
T45_2	Vehicle 2 skidding and overturning = No & Vehicle 2 manoeuvre = Going ahead & Vehicle 2 type = Car, Van & Vehicle 2 age ≤ 15	Serious	205	22.85
T26_5	Vehicle 2 manoeuvre = Going ahead & Cyclist age ≤ 17, 18–24, 25–34, 35–44, 45–54 & Bike leaving carriageway = No & Junction detail = Crossroads, Slip road	Serious	198	22.07
T36_14	Area = Urban & Driver 2 IMD = Less deprived	Serious	197	21.96
T65_1	Vehicle 2 towing and articulation = No & Vehicle 2 manoeuvre = Turning left/right/U/Reversing & Bike 1st Point of impact = Front	Serious	196	21.85
T14_5	Driver 2 Gender = M & Junction detail = No junction & Bike leaving carriageway = No	Serious	185	20.62

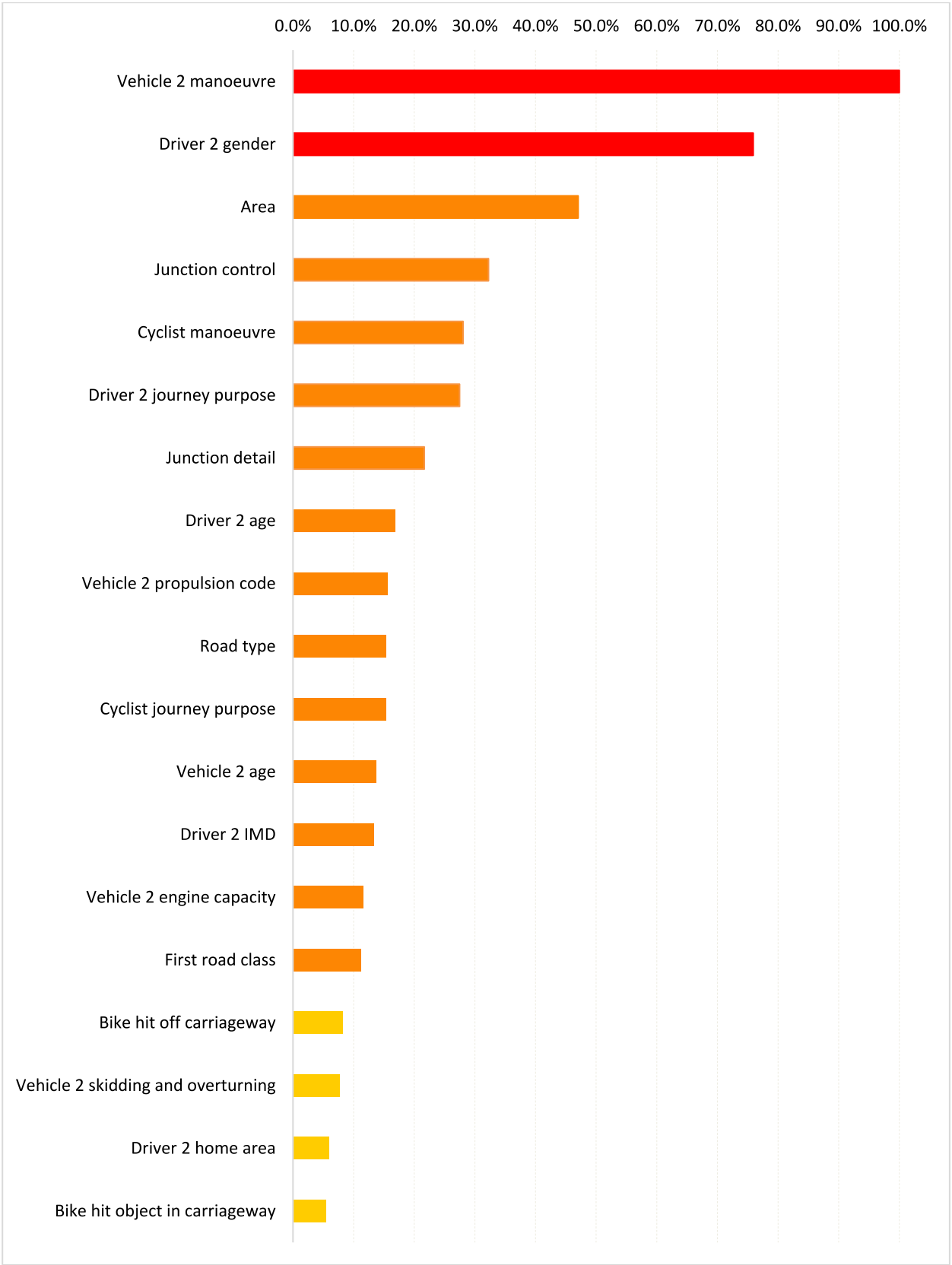

Fig. 1. Variable importance for fatal crashes.

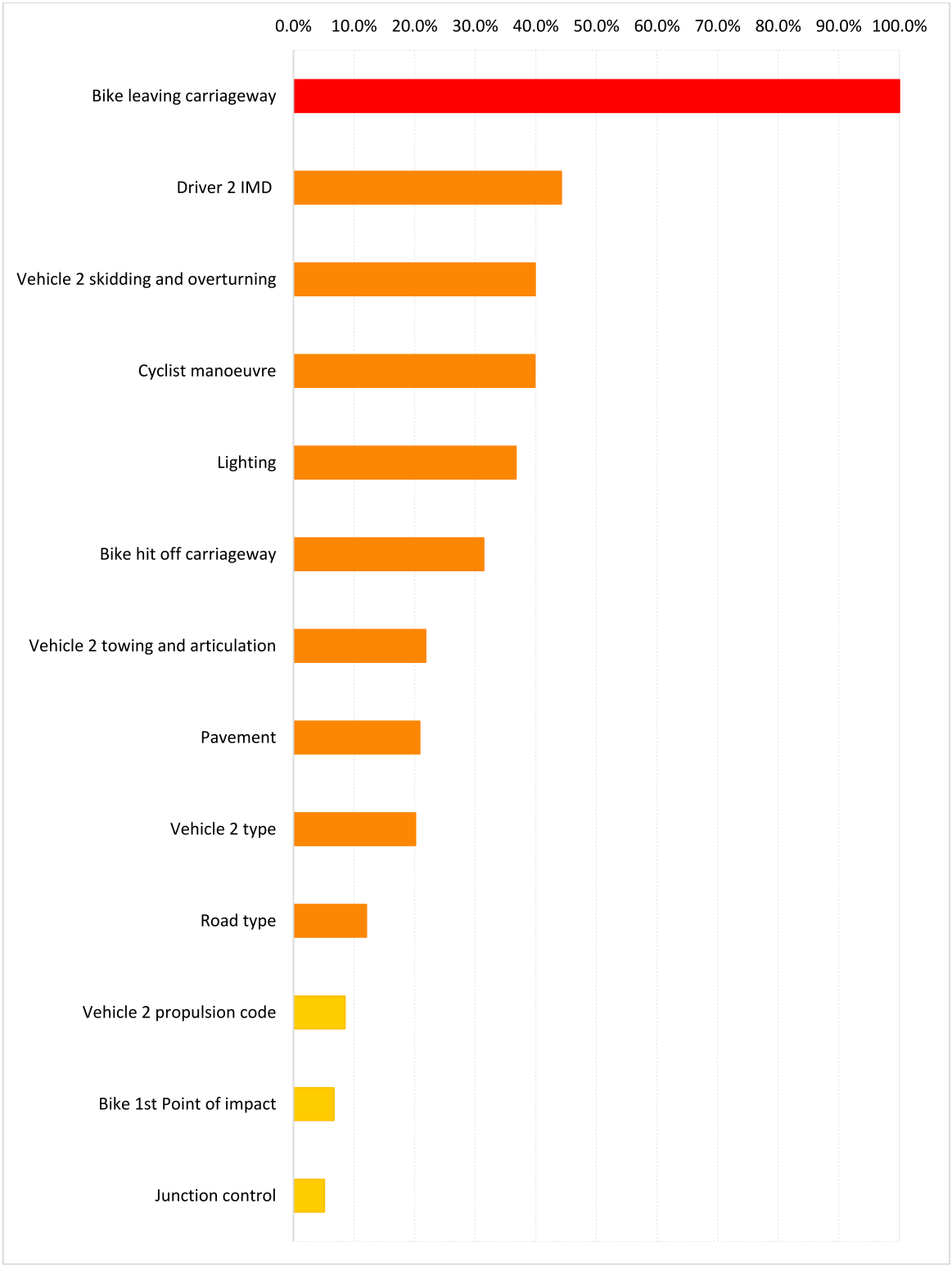

Fig. 2. Variable importance for serious cyclist crashes.

were 8 and included: (1) the IMD of the second driver involved in the serious injury cyclist crash, that showed the relative deprivation level of the area the second driver involved in the serious injury cyclist crash lives in; (2) the skidding and/or overturning of the second vehicle (yes or no); (3) cyclist manoeuvre, that included manoeuvres like going ahead, moving off, overtaking, and turning left, right, U inversion and reversing; (4) lighting, that provided information about the lighting condition at the time the crash occurred (daylight or darkness); (5) bike hit off carriageway, that provided information about the cyclist impact against barriers, poles, trees or walls when hitting off the carriageway because of the crash; (6) the presence of towed or articulated vehicles in cyclist-other vehicle crashes (yes or no); (7) pavement, that contributed to understand the pavement condition at the time of the crash occurrence (dry or wet/frozen); and (8) the type of the second vehicle

involved in the cyclist crash contributing to understand if the second vehicle involved in the crash when the outcome was a serious injury was another bike, a PTW, a car, a van, a bus or a truck. There were also 4 variables of small importance (less than 20% of importance).

The serious injury outcome was the consequent for 349 rules (1 with two antecedents, 11 with three antecedents, 337 with four antecedents, 33.6% of the total rules). As the first antecedent, the most important variable, that was the bike leaving carriageway, gave rise to 27 if-then rules. Of which, almost the half of the rules identified the cyclist not leaving the carriageway. Interestingly, the variables that provided the greatest number of rules as first item were those identified also for fatal crashes: the manoeuvre of the second vehicle involved (60 rules out of 349), the type of the second vehicle involved (54 rules out of 349), and its engine capacity (42 rules out of 349). In 37 rules, the most frequent manoeuvres associated to the second vehicle involved in the cyclist crash were turning left, turning right, U inversion, and reversing and moving off manoeuvres. The involvement of bus or truck in the cyclist crashes generated 39 rules with the serious injury outcome, followed by the PTW that was found in 34 rules. The strongest rule extracted by the RF trees was the rule T54_12 covering 50% of the observations falling in the tree from which the rule was extracted. The rule associated a cyclist that does not leave the carriageway and vehicles of different engine capacities with a serious injury crash as consequent. This rule confirmed the vulnerability of cyclists compared with motorized vehicles. The patterns of the second vehicle were further identified also affecting the serious cyclist crashes. The rule T67_14, indeed, associated the combination of manoeuvring and vehicle type with a greater probability of serious injury.

Differently from fatal crashes with cyclist involvement, the RF identified the urban area as the most frequent context in which serious injury crashes are observed. The rule T23_7 provided the association of urban area and frontal impact with serious injury crashes. Another rule (T36_14) provided the association of urban area and the driver of the second vehicle involved living in less deprived area with serious injury crashes. When involved in crashes, (rule T65_1) the vulnerability of the cyclists of being involved in serious injury outcomes increases with a frontal impact with the second vehicle involved that is making turning left or right manoeuvres, U inversion or reversing.

5.2. Random parameters logit model

In the fatal severity prediction 21 explanatory variables with 45 indicator variables were found to be statistically significant, whereas in the serious severity prediction 17 explanatory variables with 30 indicator variables proved to be statistically significant. Four normal distributed indicator variables were found to produce random parameters: cyclist age $\geq$ 75 (fatal prediction), cyclist gender male (fatal and serious prediction), and driver age 55–64 (serious prediction). The indicator variable cyclist age $\geq$ 75 was associated with a mean of 3.29, and a standard deviation of parameter density function of 0.66. This means that, for more than 99.9% of the crashes with an elderly cyclist, the probability of the fatal outcome increased, while, just for a very small percentage (less than0.1%) of the observations, the probability of a fatal outcome decreased. For fatal prediction, the mean and standard deviation of density function of cyclist gender male were respectively equal to 0.37 and − 0.97 indicating that for 64.72% of the cyclist crashes the probability of a fatal outcome increased by the presence of a male cyclist whereas, for the remaining 35.28% of the observations, presence of a male cyclist caused a decrease in that probability. For serious outcome, the mean and standard deviation of density function of cyclist gender male amounted to − 0.11 and 1.56. This implies that for 52.80% of the observations with a male cyclist the probability of the serious crash decreased, meanwhile for 47.20% of the observations the probability of a serious outcome increased. The indicator variable driver age 55–64 showed a mean of − 0.12, and a standard deviation of parameter density function of 0.51. This suggests that for 58.96% of the observations with a male driver the probability of the serious outcome decreased, whereas for 41.04% of crashes the probability of a serious outcome decreased.

The model's McFadden Pseudo R^2 is equal to 0.21, indicating a very good fit. Whereas in terms of F-measures and G-mean the model performed differently. For instance, the model ability in terms of F-measures is reasonable for serious injury (equal to 0.29) but low for fatal crashes (equal to 0.06).

Instead, G-mean exhibited highest classification performance for fatal crashes (equal to 0.77) than serious injury (equal to 0.49) (Table 12). Similarly, the FP_{rate} is lower for fatal crashes.

The results for both the fixed and random variables were shown in Table 9 (part A and part B). For each significant variable (p-value less than 0.05), the β_j estimated value was reported. Furthermore, in the Table 10 (part A and part B) the pseudo-elasticity values were provided (Table 11).

High speed limits have a significant impact on the severity of cyclist crashes. In comparison to speed limits equal to 30 mph, speed limits of 40 mph are linked to an increase of 2.67% in serious injuries and a significant increase of 19.61% in fatal crashes. The impact becomes even more pronounced for speed limits equal to or greater than 50 mph, with a substantial 8.25% higher probability of serious cyclist injuries and an alarming 36.63% higher probability of cyclist fatalities. These findings indicate a significant escalation in the likelihood of serious injuries and fatal crashes as speed limits increase.

Interestingly, in comparison to single carriageways, roundabouts exhibit a significant decrease of 22.01% in the probability of fatal cyclist crashes, making them the road type with the lower probability of such crashes. In contrast, dual carriageways and slip roads show increases of 6.41% and 13.62% respectively in the probability of cyclist fatalities. This highlights the contrasting safety profiles of different road types, with roundabouts offering a comparatively safer environment for cyclists, while dual carriageways and slip roads pose higher risks.

The presence of darkness plays a significant role in crash severity, as reflected in the pseudo-elasticity values. For fatal cyclist crashes, darkness is associated with an increase of 14.47% in the probability, indicating that cycling during darker conditions raises the risk of fatal crashes. Moreover, darkness also shows a slight increase of 2.55% in the probability of serious cyclist crashes.

Wet/frozen pavement is significant mainly for fatal crashes, with an increase of 7.75% in probability. On the other hand, weekends exhibit a broader effect, leading to an increase in both fatal and serious crashes. For fatal crashes, weekends are associated with an increase of 6.77% in the probability, suggesting a higher risk of fatal outcomes during this time. Furthermore, weekends also show a slight increase of 2.61% in the probability of serious injuries.

Bike leaving the carriageway on the nearside and offside is strongly associated with crash severity, as highlighted by the pseudo-elasticity values. When the bike leaves the carriageway on the nearside, there is a substantial 31.32% increase in the probability of fatal crashes and a noticeable 5.41% increase in the probability of serious injuries. Similarly, when the bike leaves the carriageway on the offside, the impact is even more pronounced, with a significant 65.62% increase in the probability of fatal crashes and a substantial 11.72% increase in the probability of serious injuries.

Considering vehicle 2 engine capacity with 1501–2000 as the baseline, engine capacity greater than 3000 is associated with a substantial increase of 54.26% in the probability of cyclist fatalities and a noticeable increase of 5.74% in the probability of serious injuries. These findings suggest that encounters with vehicles possessing larger engine capacities pose a significantly higher risk to cyclists in terms of both fatal outcomes and serious injuries.

Regarding driver-related factors, both male gender and young drivers ($\leq$17) increase the probability of fatal crashes by 5.52% and 4.75% respectively. The most influential variable is cyclist age. Compared to young cyclists (25–34), older cyclists exhibit an escalating risk of fatal crashes: for cyclists aged 35–44, the probability increases by

Table 9
Mixed logit: parameter estimates (Part A).

Variable	Fatal			Serious		
	β	Std. Err.	p-value	β	Std. Err.	p-value
Intercept	−0.82	0.04	<0.001	0.16	0.06	0.01
Speed Limit (30 mph as baseline)						
20	−0.13	0.02	<0.001	−0.10	0.03	<0.001
40	0.89	0.02	<0.001	0.12	0.04	<0.001
≥ 50	1.67	0.03	<0.001	0.38	0.05	<0.001
Area (Rural as baseline)						
Urban	−0.66	0.02	<0.001	−0.18	0.03	<0.001
Junction control (Not at junction or within 20 m as baseline)						
Give way/Stop	−0.74	0.01	<0.001	−0.08	0.02	<0.001
Traffic lights	0.30	0.01	<0.001			
Pedestrian crossing physical facilities (No physical crossing facilities within 50 m baseline)						
Central refuge	0.93	0.01	<0.001			
Pedestrian phase at traffic signal junction	0.44	0.02	<0.001	−0.13	0.04	<0.001
Pelican puffin toucan or similar non junction Pedestrian light crossing	0.31	0.01	<0.001			
Zebra	−0.43	0.02	<0.001			
Road type (Single carriageway as baseline)						
Dual carriageway	0.29	0.01	<0.001			
Roundabout	−1.00	0.02	<0.001			
Slip road	0.62	0.03	<0.001			
Lighting (Daylight as baseline)						
Darkness	0.66	0.02	<0.001	0.12	0.03	<0.001
Pavement (Dry as baseline)						
Wet/Frozen	0.35	0.01	<0.001			
Weather (Clear as baseline)						
Raining	−0.46	0.02	<0.001	−0.19	0.04	<0.001
Day of week (Weekday as baseline)						
Weekend	0.31	0.01	<0.001	0.12	0.03	<0.001
Number of bikes (1 as baseline)						
>1	1.04	0.05	<0.001	0.84	0.07	<0.001
Bike 1st point of impact (No impact as baseline)						
Back				−0.58	0.05	<0.001
Front	−0.68	0.02	<0.001	−0.16	0.04	<0.001
Nearside/Offside	−0.66	0.03	<0.001	−0.34	0.04	<0.001
Bike leaving carriageway (No as baseline)						
Nearside	1.43	0.04	<0.001	0.25	0.06	<0.001
Offside	2.99	0.08	<0.001	0.54	0.13	<0.001
Vehicle 2 skidding and overturning (No as baseline)						
Yes	0.37	0.03	<0.001	0.40	0.02	<0.001
Vehicle 2 engine capacity (1501–2000 as baseline)						
≤ 1000	0.50	0.01	<0.001			
2001–3000	0.08	0.02	<0.001			
>3000	2.47	0.04	<0.001	0.26	0.06	<0.001
Vehicle 2 age (≤15 as baseline)						
>15				−0.49	0.01	<0.001
Table 9. Mixed logit: parameter estimates (Part B)						
Variable	**Fatal**	**Serious**				
	β	**Std. Err.**	**p-value**	**β**	**Std. Err.**	**p-value**
Intercept	−0.82	0.04	<0.001	0.16	0.06	0.01
Driver 2 gender (Female as baseline)						
Male	0.25	0.01	<0.001			
Driver 2 age (25–44 as baseline)						
≤17	0.22	0.06	<0.001			
18–24	−0.24	0.02	<0.001			
35–44	−0.41	0.02	<0.001	−0.21	0.03	<0.001
45–54	−0.73	0.02	<0.001	−0.14	0.04	<0.001
55–64	−0.19	0.02	<0.001	−0.12	0.05	0.01
St. dev. of density function 55–64	0.51	0.19	0.01			
65–74	−1.35	0.03	<0.001	−0.12	0.05	0.02
≥75	−0.94	0.03	<0.001	0.13	0.06	0.03
Cyclist IMD (Less deprived as baseline)						
More deprived	0.14	0.01	<0.001			
Cyclist gender (Female as baseline)						
Male	0.37	0.02	<0.001	−0.11	0.03	<0.001
St. dev. of density function Male	−0.97	0.03	<0.001	1.56	0.06	<0.001
Cyclist age (25–34 as baseline)						
≤17	−0.07	0.02	<0.001			
18–24	0.12	0.02	<0.001			
35–44	0.43	0.02	<0.001	0.28	0.03	<0.001
45–54	0.61	0.02	<0.001	0.47	0.03	<0.001
55–64	1.40	0.02	<0.001	0.6	0.04	<0.001
65–74	2.35	0.04	<0.001	0.69	0.06	<0.001
≥75	3.29	0.07	<0.001	0.74	0.10	<0.001
St. dev. of density function ≥ 75	−0.66	0.10	<0.001			
Cyclist journey purpose (Commuting to from work as baseline)						
Journey as part of work	−0.39	0.03	<0.001			
To/from school	0.30	0.03	<0.001	−0.27	0.07	<0.001
Number of observations	72,363					
Log likelihood null model	−79,499					
Log likelihood full model	−62,668					

IMD = Index of Multiple Deprivation.

Table 10
Pseudo-elasticity for significant variables (Part A).

Variable	Fatal	Serious
Speed Limit (30 mph as baseline)		
20	−2.96%	−2.23%
40	19.61%	2.67%
≥ 50	36.63%	8.25%
Area (Rural as baseline)		
Urban	−14.65%	−4.03%
Junction control (Not at junction or within 20 m as baseline)		
Give way/Stop	−16.34%	−1.72%
Traffic lights	6.49%	
Pedestrian crossing physical facilities (No physical crossing facilities within 50 m baseline)		
Central refuge	20.36%	
Pedestrian phase at traffic signal junction	9.74%	−2.85%
Pelican puffin toucan or similar non junction Pedestrian light crossing	6.78%	
Zebra	−9.44%	
Road type (Single carriageway as baseline)		
Dual carriageway	6.41%	
Roundabout	−22.01%	
Slip road	13.62%	
Lighting (Daylight as baseline)		
Darkness	14.47%	2.55%
Pavement (Dry as baseline)		
Wet/Frozen	7.75%	
Weather (Clear as baseline)		
Raining	−10.15%	−4.21%
Day of week (Weekday as baseline)		
Weekend	6.77%	2.61%
Number of bikes (1 as baseline)		
>1	22.87%	18.29%
Bike 1st point of impact (No as baseline)		
Back		−12.63%
Front	−14.87%	−3.60%
Nearside/Offside	−14.50%	−7.52%
Bike leaving carriageway (No as baseline)		
Nearside	31.32%	5.41%
Offside	65.62%	11.72%
Vehicle 2 skidding and overturning (No as baseline)		
Yes	8.68%	8.00%
Vehicle 2 engine capacity (1501–2000 as baseline)		
≤ 1000	10.91%	
2001–3000	1.78%	
>3000	54.26%	5.74%
Vehicle 2 age (≤15 as baseline)		
>15		−10.67%

Table 11
Pseudo-elasticity for significant variables (Part B).

Variable	Fatal	Serious
Driver 2 gender (Female as baseline)		
Male	5.52%	
Driver 2 age (25–44 as baseline)		
≤17	4.75%	
18–24	−5.23%	
35–44	−8.99%	−4.58%
45–54	−15.92%	−3.00%
55–64	−4.13%	−2.52%
65–74	−29.55%	−2.61%
≥75	−20.60%	2.75%
Cyclist IMD (Less deprived as baseline)		
More deprived	3.01%	
Cyclist gender (Female as baseline)		
Male	8.06%	−2.40%
Cyclist age (25–34 as baseline)		
≤17	−1.67%	
18–24	2.59%	
35–44	9.55%	6.14%
45–54	13.47%	10.26%
55–64	30.67%	13.01%
65–74	51.51%	15.15%
≥75	73.79%	16.44%
Cyclist journey purpose (Commuting to from work as baseline)		
Journey as part of work	−8.44%	
To/from school	6.65%	−5.87%

Table 12
Performance measures of the RPLM.

Severity level	F-measure	G-mean	FP_{rate}	McFadden R^2
Fatal	0.06	0.77	0.15	0.21
Serious injury	0.29	0.49	0.24	
Slight injury	0.72	0.53	0.43	

9.55%. It further rises to 13.47% for cyclists aged 45–54, 30.67% for cyclists aged 55–64, 51.51% for cyclists aged 65–74, and a significant 73.79% for cyclists aged ≥ 75. These findings highlight the increasing risk associated with older age groups in terms of fatal outcomes. Thus, as cyclist age increases, the probability of experiencing a fatal or a serious crash significantly rises.

Finally, the purpose of the cyclist's journey, specifically when it is to or from school, influences the probability of fatal crashes, as indicated by the pseudo-elasticity value of 6.65%. When the cyclist's journey is specifically related to commuting to or from school, there is a notable 6.65% increase in the probability of fatal outcomes.

6. Discussion

Study results identified several roadways, environmental, vehicle, driver, and cyclist-related factors associated with cyclist crash severity. The integrated use of the RF and the RPLM yielded significant findings in analysing the cyclist crash severity. Both RF and RPLM models found that rural area, characterized by higher speed limits, was where the most severe crashes occurred. Indeed, the probability of a serious or a fatal outcome increases exponentially with the speed limit. This finding is consistent with previous studies (Cloutier at al., 2019; Isaksson-Hellman and Töreki, 2019). The higher the speed, the greater the stopping distance required, and hence the increased probability to have the most severe injury. To mitigate these effects, we recommend specific safety countermeasures such as traffic calming and low speed zones in areas with significant cyclists' activity, and strict speed enforcement as carried out in Oslo and Helsinki that achieved the zero pedestrian and cycling fatality goal by decreasing the traffic speeds (ETSC, 2020).

Regarding the road related variables, the results of both models further showed that the likelihood of fatal crash occurrence increases on dual carriageways, slip roads and at junctions controlled by traffic lights. Although several studies have identified intersections as dangerous areas (Moore et al., 2011; Yuan and Abdel-Aty, 2018), the RPLM showed that the probability of a fatal crash in the roundabout is lower than on dual or single carriageways. This is mainly due to the reduction of vehicle speeds in roundabouts (Montella, 2011, 2018).

As regards the light condition, nighttime is related to the increase of injury severity. The likelihood of fatal and serious crash occurrence at night is much higher than during the daytime. Cyclist visibility is an important road safety issue, especially in unlit streets at night where both drivers and cyclist sight are critically reduced (Chen and Funny, 2019; Hu et al., 2022). Lighting with light emitting diodes (LEDs), mandatory bicycle lights, and reflective clothing for cyclists can improve visibility during the nighttime.

Regarding the cyclist characteristics, male cyclists were found to have an increased likelihood of fatal outcome because of their higher tendency for risky behaviour. This is in line with the findings of previous research (Bíl et al., 2010; Hosseinpour et al., 2021; Eluru et al., 2008; Meredith et al., 2020). The age of the cyclist was also found to be a statistically significant variable, with elderly cyclists (age ≥ 75) showing a dramatic increase in the probability of a fatal outcome. This result might be due to the greater fragility of older cyclists as well as to their increased reaction and perception times. These observations are in line with the findings of previous research (Bahrololoom et al., 2020; Liu et al., 2020; Oikawa et al., 2019; Samerei et al., 2021). Safety

improvement strategies focused on male and elderly cyclists should include education and/or training of cyclists through programs aimed at teaching safe cycling skills, the building of buffered cycle lanes with forgiving curbs to consider that people make mistakes, the removal of obstacles from the bike path, and regulations requiring the use of helmet while cycling. Helmet use has been shown to be 85 and 88% effective in mitigating head and brain injuries while nearly 70% of bicyclist fatalities involve head injuries (WHO, 2020). The increasing helmet use is essential for future improvements in cyclist safety. Moreover, the bike helmets with airbags developed in Sweden produced a significantly reduction in the risk of most severe injuries (WHO, 2020).

Although "poverty" and "deprivation" have often been considered as synonyms, there is a clear distinction between the two terms. Deprivation refers to people's unmet needs whereas poverty refers to the lack of resources required to meet those needs (McLennan et al., 2019). Thus, people are deprived if they lack the resources of all kinds, not just income. The main measurement of deprivation for England is the IMD, which is a measure of relative deprivation of small areas roughly equivalent to postcode areas, each with a similar population size. It is based on 7 different domains of deprivation (income deprivation, employment deprivation, education, skills and training deprivation, health deprivation and disability, crime, barriers to housing and services and living environment deprivation), defining deprivation in a broad way to encompass a wide range of aspects of an individual's living conditions (Townsend, 1979). Since the Department of Transport (2022b) found an association between casualties and deprivation at the national level, we investigated driver and cyclist IMD. The mixed logit model showed that cyclists from more deprived areas were more likely to die due to a crash. Probably, this is for the lack of adequate cyclist facility and safety education. Thus, it's necessary to invest equitable among different areas in well-connected and continue cycling infrastructure and to develop educational programs focused on road safety.

The RF model identified the gender of the second vehicle' driver as one of the variables having the greatest normalized importance and higher predictive ability for fatal classification. Like the effects of male cyclist, the mixed logit found male drivers increasing the likelihood of fatal crash relative to their female counterparts. Young drivers were also more likely to be involved in the fatal crashes, and results of both econometric and machine learning methods confirm previous findings (Scholes et al., 2018). These outcomes may be explained by the poor skills of young drivers and the driving style of young or male drivers that is usually more aggressive. The implementation and exercising of legal regulations concerning road safety, targeted safety campaigns, and educational programs for young and male drivers may contribute to the change of bed behaviours and the increase in awareness.

Regarding the second vehicle involved in the cyclist crash characteristics, RF model has associated vehicles older than 15 years with an increasing in likelihood of fatal crashes, while mixed logit model has identified older vehicles as a predictor of serious crashes. Although previous studies showed that increasing vehicle age increases the probability of most severe injuries (Behnood and Mannering, 2015), the association between vehicle age and cyclist crash severity is complex (Behnood et al., 2014). As the vehicle type, bus and truck occurred in 56 rules with fatal crash as consequent and in 39 rules with the serious injury outcome. These results are in line with the findings of previous research (Damsere-Derry and Bawa, 2018; Joo et al., 2017; Mason-Jones et al., 2022; Sun et al., 2022a).

The overall findings regarding the vehicle type are due to the larger mass and the wider stiffness, the major area of impact for cyclists, the higher bumper height, and the greater stopping distances that characterized bus and truck as compared with other vehicles. In London, even though heavy goods vehicles account just for a very small percentage of vehicle kilometres, they are involved in 20% of the cyclist fatal crashes (RoSPA, 2015). The truck side guards, also known as lateral protective devices (designed to keep pedestrians and cyclists from being run over by rear wheels of a large truck), may be effective in reducing the severity of crashes between trucks and cyclists. Overpasses and underpasses that enable cyclists to cross distributor roads without sharing the space with motorized vehicles, separate facilities for vehicle types (trucks/bus and bikes), mirrors on signal posts to enable large vehicle drivers to see in otherwise blind spots, the use of safer trucks with a better visibility and education of drivers, cyclists and road managers could be also useful measures to improve cyclist safety. Both econometric and machine learning methods showed that higher engine capacity of second vehicle involved in the crash increases the probability of fatal outcome. This is likely because, the motor vehicle with a larger engine capacity tends to attain higher speed and such drivers are also more likely to be aggressive assuming a risk behaviour (Mazharul Haque et al., 2009).

As regards the journey purpose, the models identified the most dangerous situation as the way to and from school, situations when the children are unprotected road users (European Commission, 2012). Furthermore, the RF model found the variable vehicle 2 manoeuvre having the greatest normalized importance for fatal classification. The RF rules showed that the most frequent manoeuvres associated to the second vehicle involved in the fatal crash were the turning left, turning right, U inversion, and reversing. These results are in line with previous studies (Behnood and Mannering, 2017; van Haperen et al., 2018, Wang et al., 2018) and suggest planning specific protected cycle lanes and providing separation for turning manoeuvres of cyclists and motor vehicles. Other relevant countermeasures for crash severity reduction include Intelligent Transportation System (ITS) safety applications that have proved their effectiveness reducing road fatalities by minimizing human error. The ITS applications for enhancing cyclist safety involve road user presence warning signals, traffic signal prioritisation based on the detection of cyclists, intelligent cycling infrastructure which reflect patterns of cycle flows within defined areas (such as major destinations – schools, train stations or university), and Advanced Driver Assistance Systems (ADAS). Most of the existing ADAS are designed for motorized vehicles (e.g., autonomous emergency braking,). However, they can also be available for cyclists, either as an application on a smartphone or as a dedicated device, which can be integrated in the bike (Scarano et al., 2023; Scholliers et al, 2017).

7. Conclusions

In this paper, both a machine learning method, as the RF, and an econometric model, as the RPLM, were implemented to identify several patterns associated with severe injury and fatal cyclists' crashes. These associations were examined using crash data from Great Britain in the period between 2016 and 2018.

The Random Survival Forest demonstrated higher prediction accuracy among the three performed RF algorithms. The RF tool showed that the interaction between a bike and other motorized vehicles is likely to increase the crash severity. Such critical interactions were only identified by the RF. Moreover, the second vehicle manoeuvre, the second vehicle type, the driver's gender, journey purpose, and IMD were also identified by the RF variable importance rank as the variables exhibiting the highest detrimental impact on crash severity.

The RPLM showed that four indicator variables, such as cyclist age $\geq$ 75 (fatal prediction), cyclist gender male (fatal and serious prediction), and driver age 55–64 (serious prediction), were associated with normally distributed random parameters, with statistically significant standard deviation of parameter density function indicating significant unobserved heterogeneity in the data. Furthermore, the pseudo-elasticity of the explanatory variables provided the magnitude of the effects of each variable on injury severity probabilities, gaining valuable insights into the relative importance and influence of the variables. For instance, speed limit emerged as a significant factor, with higher limits demonstrating positive elasticities for both fatal and serious crashes. The junction control and road characteristics also showed notable effects. Specifically, the presence of zebra crossings and roundabouts are associated with a negative pseudo-elasticity, indicating their potential as

safety-enhancing features. Lighting conditions as well as the number of bikes involved in the crash played a significant role, with the darkness and the presence of multiple bikes involved in the crash increase the probabilities of experiencing fatal and serious outcome. While the RPLM felt short in capturing the effects of the second vehicle, the RF provided several results related to the effects of the second vehicle on crash severity.

Finally, focusing specifically on factors related to the cyclist, the pseudo-elasticities values revealed that a cyclist leaving the carriageway on the nearside or offside as well as the older age cyclist groups (55–64, 65–74, ≥75) strengthened the possibility of both fatal and serious outcome. As cyclist age increases, the probability of experiencing a fatal or a serious crash significantly rises.

The RPLM identified a larger number of predictors compared to the RF. However, the RF method automatically detected patterns and relationships within datasets being able to uncover hidden correlations that were not captured by the econometric models. The insights provided by the machine learning tool are valuable because they shed light on the interdependence among the different factors related to road infrastructure, the environment, the vehicle characteristics, and the driver and cyclist influencing cyclist' crashes. This interdependence refers to how these factors interact leading to the severity of crashes.

As regards the methodological perspective, this study shows that safety analyses combining both econometric and machine learning models are very useful and informative. Econometric models may provide quantitative and easy interpretable insights into the factors influencing cyclist crash severity, while machine learning models may detect complex crash scenarios and uncover hidden correlations among the contributing factors. By integrating these two approaches, policymakers and researchers can gain a comprehensive understanding of the factors contributing to cyclist safety and develop targeted interventions and policies to mitigate risks and improve overall safety for cyclists.

In this study, the random parameters logit model was estimated without accounting for heterogeneity in means and variances. A correlated and grouped random parameters model with heterogeneity in means and/or variances is thus the direction of future research.

CRediT authorship contribution statement

Antonella Scarano: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Maria Rella Riccardi:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Filomena Mauriello:** Conceptualization, Methodology, Software, Writing – review & editing. **Carmelo D'Agostino:** Conceptualization, Writing – review & editing. **Nicola Pasquino:** Conceptualization, Methodology, Writing – review & editing. **Alfonso Montella:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

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

Data availability

Available online

References

Ahmed, S.S., Pantangi, S.S., Eker, U., Fountas, G., Still, S.E., Anastasopoulos, P.C., 2020. Analysis of safety benefits and security concerns from the use of autonomous vehicles: A grouped random parameters bivariate probit approach with heterogeneity in means. Anal. Method. Accid. Res. 28, 100134.

Ahmed, S.S., Cohen, J., Anastasopoulos, P.C., 2021. A correlated random parameter with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. Anal. Method. Accid. Res. 30, 100160 https://doi.org/10.1016/j.amar.2021.100160.

Ahmed, S.S., Fountas, G., Anastasopoulos, P.C., Peeta, S., 2023. Analysis of urban travel time and travel distance: A fully parametric bivariate hazard-based duration modelling approach with correlated grouped random parameters. Travel Behav. Soc. 31, 271–283.

Akgun, N., Dissanayake, D., Thorpe, N., Bell, M.C., 2018. Cyclist casualty severity at roundabouts - to what extent do the geometric characteristics of roundabouts play a part? J. Saf. Res. 67, 83–91.

Alshehri, A., Eustace, D., Hovey, P., 2020. Analysis of factors affecting crash severity of pedestrian and bicycle crashes involving vehicles at intersections. International Conference on Transportation and Development 2020 -Traffic and Bike/Pedestrian Operations.

Amaratunga, D., Cabrera, J., Lee, Y.S., 2008. Enriched random forests. Bioinformatics, 24 (18), 2010-2014. doi: 10.1093/bioinformatics/btn356.

Anastasopoulos, P.C., Mannering, F., 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. Accid. Anal. Prev. 43 (3), 1140–1147. https://doi.org/10.1016/j.aap.2010.12.024.

Andreß, H.-J., Golsch, K., Schmidt, A.W. (Eds.), 2013. Applied Panel Data Analysis for Economic and Social Surveys. Springer Berlin Heidelberg, Berlin, Heidelberg.

Asgarzadeh, M., Fischer, D., Verma, S.K., Courtney, T.K., Christiani, D.C., 2018. The impact of weather, road surface, time-of-day, and light conditions on severity of bicycle-motor vehicle crash injuries. Am. J. Ind. Med. 61 (7), 556–565.

Bahrololoom, S., Young, W., Logan, D., 2020. Modelling injury severity of bicyclists in bicycle-car crashes at intersections. Accid. Anal. Prev. 144, 105597.

Bai, L., Sze, N.N., 2020. Red light running behavior of bicyclists in urban area: Effects of bicycle type and bicycle group size. Travel Behav. Soc. 21, 226–234. https://doi.org/10.1016/j.tbs.2020.07.003.

Behnood, A., and Mannering, F., 2017. Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variance. Anal. Method. Accid. Res., 16, 35-47, Doi: 10.1016/j.amar.2017.08.001.

Behnood, A., Mannering, F., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. Anal. Method. Accid. Res. 8, 7–32. https://doi.org/10.1016/j.amar.2014.10.001.

Behnood, A., Roshandeh, A., Mannering, F., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Anal. Method. Accid. Res. 3–4, 56–91. https://doi.org/10.1016/j.amar.2015.08.001.

Bekkar, M., Djemaa, H.K., Alitouche, T.A., 2013. Evaluation Measures for Models Assessment over Imbalanced Data Sets. J. Informat. Eng. Appl. 3(10), ISSN 2225-0506.

Bíl, M., Bílová, M., Müller, I., 2010. Critical factors in fatal collisions of adult cyclists with automobiles. Accid. Anal. Prev. 42 (6), 1632–1636. https://doi.org/10.1016/j.aap.2010.04.001.

Blaizot, S., Papon, F., Haddak, M.m., Amoros, E., 2013. Injury incidence rates of cyclists compared to pedestrians, car occupants and powered two-wheeler riders, using a medical registry and mobility data, rhone county, france. Accid. Anal. Prev. 58 https://doi.org/10.1016/j.aap.2013.04.018.

Boufous, S., de Rome, L., Senserrick, T., Ivers, R., 2012. Risk factors for severe injury in cyclists involved in traffic crashes in victoria, australia. Accid. Anal. Prev. 49, 404–409.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees; Wadsworth International Group: Belmont, CA, USA. doi: 10.1201/9781315139470.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5–32. Kluwer Academic Publishers. Manufactured in The Netherlands. Available online: https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf.

Buhler, T., Comby, E., Vaudor, L., von Pape, T., 2021. Beyond 'good' and 'bad' cyclists. On compensation effects between risk taking, safety equipment and secondary tasks. J. Transp. Health 22 (101131). https://doi.org/10.1016/j.jth.2021.101131.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accid. Anal. Prev. 42 (4), 1072–1079. https://doi.org/10.1016/j.aap.2009.12.015.

Calvi, A., D'Amico, F., Ferrante, C., Bianchini Ciampoli, L., 2021. Driving Simulator Study for Evaluating the Effectiveness of Virtual Warnings to Improve the Safety of Interaction Between Cyclists and Vehicles. Transp. Res. Rec. 2676 (4), 436–447. https://doi.org/10.1177/03611981211061351.

Carlson, M.B., Barbour, M., Abdel-Aty, M., 2023. Effectiveness of bicycle helmets and injury prevention: a systematic review of meta-analyses. Sci. Rep. 13 (1), 8540. https://doi.org/10.1038/s41598-023-35728-x.

Chang, F., Haque, M.M., Yasmin, S., Huang, H., 2022. Crash injury severity analysis of E-Bike Riders: A random parameters generalized ordered probit model with heterogeneity in means. Saf. Sci. 146, 105545 https://doi.org/10.1016/j.ssci.2021.105545.

Chen, H., and Funny, K., 2019. Understanding the Contributing Factors to Nighttime Crashes at Freeway Mainline Segments. J. Transport. Technol., 9, 450-461, https://www.scirp.org/journal/jtts .

Chen, P., Shen, Q., 2016. Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. Accid. Anal. Prev. 86, 239–246. https://doi.org/10.1016/j.aap.2015.11.002.

Damsere-Derry, J., Bawa, S., 2018. Bicyclists' accident pattern in northern ghana. Iatss Research 42 (3), 138–142.

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., Jalayer, M., 2019. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for

appropriate countermeasures. Int. J. Urban Sci. 23, 38–40. https://doi.org/10.1080/12265934.2018.1431146.
Dash, I., Abkowitz, M., Philip, C., 2022. Factors impacting bike crash severity in urban areas. J. Saf. Res. 83, 128–138. https://doi.org/10.1016/j.jsr.2022.08.010.
DfT – Department for Transport, 2022b. Reported road casualties in Great Britain: Casualties and deprivation.
DfT – Department for Transport, 2022a. Reported road casualties in Great Britain: pedal cycle factsheet, 2021.
Du, W., Yang, J., Powis, B., Zheng, X., Ozanne-Smith, J., Bilston, L., Wu, M., 2013. Understanding on-road practices of electric bike riders: an observational study in a developed city of China. Accid. Anal. Prev. 59, 319–326. https://doi.org/10.1016/j.aap.2013.06.011.
Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40 (3), 1033–1054. https://doi.org/10.1016/j.aap.2007.11.010.
European Transport Safety Council (ETSC), 2020. Zero cyclist and pedestrian deaths in Helsinki and Oslo last year. Available at https://etsc.eu/zero-cyclist-and-pedestrian-deaths-in-helsinki-and-oslo-last-year/.
European Commission, 2012. Final Report Summary - SAFEWAY2SCHOOL (Integrated system for safe transportation of children to school).
Fountas, G., Sarwar, M.T., Anastasopoulos, P.C., Blatt, A., Majka, K., 2018. Analysis of stationary and dynamic factors affecting highway accident occurrence: A dynamic correlated grouped random parameters binary logit approach. Accid. Anal. Prev. 113, 330–340.
Ghomi, H., Bagheri, M., Fu, L., Miranda-Moreno, L.F., 2016. Analyzing injury severity factors at highway railway grade crossing accidents involving vulnerable road users: a comparative study. Traffic Inj. Prev. 17 (8), 833–841. https://doi.org/10.1080/15389588.2016.1151011.
Gitelman, V., Korchatov, A., 2022. Safety-related behaviours of e-cyclists on urban streets: an observational study in Israel. Transp. Res. Procedia 60, 609–616.
Greene, W.H., Hensher, D.A., Rose, J., 2006. Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. Transp. Res. B 40, 75–92. https://doi.org/10.1016/j.trb.2005.01.005.
Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G., 2008. On the Class Imbalance Problem. Fourth International Conference on Natural Computation, 4, 192–201. doi: 10.1109/ICNC.2008.871.
Guo, Y., Li, Z., Wu, Y., Xu, C., 2018. Exploring unobserved heterogeneity in bicyclists' red-light running behaviors at different crossing facilities. Accid. Anal. Prev. 115, 118–127. https://doi.org/10.1016/j.aap.2018.03.006.
Hamann, J.C., Peek-Asa, C., Lynch, C.F., Ramirez, M., Hanley, P., 2015. Epidemiology and spatial examination of bicycle-motor vehicle crashes in Iowa, 2001–2011. J. Transp. Health 2 (2), 178–188. https://doi.org/10.1016/j.jth.2014.08.006.
Hosseinpour, M., Madsen, T.K.O., Olesen, A.V., Lahrmann, H., 2021. An in-depth analysis of self-reported cycling injuries in single and multiparty bicycle crashes in Denmark. J. Saf. Res. 77, 114–124. https://doi.org/10.1016/j.jsr.2021.02.009.
Hu, L., Wu, X., Hu, X., Wang, F., Wu, N.,2022. Injury severity analysis of electric bike crashes in Changsha, Hunan Province: taking different lighting conditions into consideration. Transport. Saf. Environ., 4 (3), doi: 10.1093/tse/tdac011.
Isaksson-Hellman, I., Töreki, J., 2019. The effect of speed limit reductions in urban areas on cyclists' injuries in collisions with cars. Traffic Inj. Prev. 20 (sup3), 39–44.
Ishwaran, H., Kogalur, U.B., 2007. Random Survival Forests for R. R News, 7(2), ISSN 1609-3631. Available at: https://www.ishwaran.org/papers/randomSurvivalForests.pdf.
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random Survival Forests. Ann. Appl. Stat. 2 (3), 841–860. https://doi.org/10.1214/08-AOAS169.
Islam, Z., Abdel-Aty, M., 2023. Traffic conflict prediction using connected vehicle data. Anal. Method. Accid. Res. 39, 100275 https://doi.org/10.1016/j.amar.2023.100275.
Jahangiri, A., Elhenawy, M., Rakha, H., Dingus, T.A., 2016. Investigating cyclist violations at signal-controlled intersections using naturalistic cycling data. In: IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio De Janeiro. https://doi.org/10.1109/ITSC.2016.7795977.
Kaplan, S., Prato, C.G., 2013. Cyclist-Motorist Crash Patterns in Denmark: A Latent Class Clustering Approach. Traffic Inj. Prev. 14 (7), 725–733. https://doi.org/10.1080/15389588.2012.759654.
Kaplan, S., Vavatsoulas, K., Prato, C.G., 2014. Aggravating and mitigating factors associated with cyclist injury severity in Denmark. J. Saf. Res. 50, 75–82.
Kent, T., Miller, J., Shreve, C., Allenback, G., Wentz, B., 2021. Comparison of injuries among motorcycle, moped and bicycle traffic accident victims. Traffic Inj. Prev. 23 (1), 34–39.
Klassen, J., El-Basyouny, K., Islam, M.T., 2014. Analyzing the severity of bicycle-motor vehicle collision using spatial mixed logit models: A city of edmonton case study. Saf. Sci. 62, 295–304.
Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M., Rakotonirainy, A., Chen, F., 2021. Crash severity analysis of vulnerable road users using machine learning. PLoS One 16 (8), e0255828.
Lapparent, M., 2005. Individual cyclists' probability distributions of severe/fatal crashes in large French urban areas. Accid. Anal. Prev. 37, 1086–1092. https://doi.org/10.1016/j.aap.2005.06.006.
Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371 https://doi.org/10.1016/j.aap.2019.105371.
Lin, Z., Fan, W., 2019. Modeling bicyclist injury severity in bicycle-motor vehicle crashes that occurred in urban and rural areas: a mixed logit analysis. Can. J. Civ. Eng. 46 (10), 924–933.
Lin, Z., Fan, W., 2021. Cyclist injury severity analysis with mixed-logit models at intersections and nonintersection locations. J. Transport. Safet. Secur. 13 (2), 223–245.
Liu, S., Fan, W., 2021. Investigating factors affecting injury severity in bicycle-vehicle crashes: a day-of-week analysis with partial proportional odds logit models. Can. J. Civ. Eng. 48 (8), 941–947.
Liu, J., Khattak, A.J., Li, X., Nie, Q., Ling, Z., 2020. Bicyclist injury severity in traffic crashes: a spatial approach for geo-referenced crash data to uncover non-stationary correlates. J. Saf. Res. 73, 25–35.
Loo, B., Tsui, K., 2010. Bicycle crash casualties in a highly motorized city. Accid. Anal. Prev. 42 (6), 1902–1907. https://doi.org/10.1016/j.aap.2010.05.011.
López, G., Abellán, J., Montella, A., de Oña, J., 2014. Patterns of Single-Vehicle Crashes on Two-Lane Rural Highways in Granada Province, Spain: In-Depth Analysis through Decision Rules. Transp. Res. Rec. 2432, 133–141. https://doi.org/10.3141/2432-16.
Lord, D., Qin, X., Geedipally, S.R., 2021. Highway Safety Analytics And Modeling. ISBN: 978-0-12-816818-9.
Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Method. Accid. Res. 25, 100113 https://doi.org/10.1016/j.amar.2020.100113.
Mason-Jones, A.J., Turrell, S., Gomez, G.Z., Tait, C., Lovelace, R., 2022. Severe and Fatal Cycling Crash Injury in Britain: Time to Make Urban Cycling Safer. J. Urban Healt 99, 334–343. https://doi.org/10.1007/s11524-022-00617-7.
McFadden, D., Train, K.E., 2000. Mixed MNL models for discrete response. J. Appl. Economet. 15, 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5%3C447::AID-JAE570%3E3.0.CO;2-1.
McLennan, D., Noble, S., Noble, M., Plunkett, E., Wright, G., Gutacker, N., 2019. The English Indices of Deprivation 2019. Technical report. London: Ministry of Housing, Communities and Local Government, https://dera.ioe.ac.uk/ 34259/1/IoD2019_Technical_Report.pdf.
Meredith, L., Kovaceva, J., Bálint, A., 2020. Mapping fractures from traffic accidents in Sweden: how do cyclists compare to other road users? Traffic Inj. Prev. 21 (3), 209–214.
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 (4), 1451–1463. https://doi.org/10.1016/j.aap.2011.02.023.
Montella, A., 2018. Roundabouts. Transport Sustainab. 11, 147–174. https://doi.org/10.1108/S2044-994120180000011009.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2011. Data-Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transp. Res. Rec. 2237, 107–116. https://doi.org/10.3141/2237-12.
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. https://doi.org/10.1016/j.aap.2011.04.025.
Montella, A., de Oña, R., Mauriello, F., Rella Riccardi, M., Silvestro, G., 2020. A data mining approach to investigate patterns of powered two-wheeler crashes in Spain. Accid. Anal. Prev. 134, 105251 https://doi.org/10.1016/j.aap.2019.07.027.
Montella, A., Imbriani, L.L., 2015. Safety performance functions incorporating design consistency variables. Accid. Anal. Prev. 74, 133–144. https://doi.org/10.1016/j.aap.2014.10.019.
Montella, A., Mauriello, F., Pernetti, M., Rella Riccardi, M., 2021. Rule discovery to identify patterns contributing to overrepresentation and severity of run-off-the-road crashes. Accid. Anal. Prev. 155, 106119 https://doi.org/10.1016/j.aap.2021.106119.
Moore, D.N., Schneider, W.H., Savolainen, P.T., Farzaneh, M., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accid. Anal. Prev. 43 (3), 621–630. https://doi.org/10.1016/j.aap.2010.09.015.
Moral-Garcia, S., Castellano, J.G., Mantas, J.G., Montella, A., Abellan, J., 2019. Decision tree ensemble method for analyzing traffic accidents of novice drivers in urban areas. Entropy 21, 360. https://doi.org/10.3390/e21040360.
Nilsson, P., Stigson, H., Ohlin, M., Strandroth, J., 2017. Modelling the effect on injuries and fatalities when changing mode of transport from car to bicycle. Accid. Anal. Prev. 100, 30–36. https://doi.org/10.1016/j.aap.2016.12.020.
Oikawa, S., Matsui, Y., Nakadate, H., Aomura, S., 2019. Factors in fatal injuries to cyclists impacted by five types of vehicles. Int. J. Automot. Technol. 20 (1), 197–205.
Ouni, F., Belloumi, M., 2018. Spatio-Temporal pattern of vulnerable road user's collisions hot spots and related risk factors for injury severity in tunisia. Transport. Res. Part F-Traffic Psychol. Behav. 56, 477–495.
Pantangi, S.S., Fountas, G., Anastasopoulos, P.C., Pierowicz, J., Majka, K., Blatt, A., 2020. Do High Visibility Enforcement programs affect aggressive driving behavior? An empirical analysis using Naturalistic Driving Study data. Accid. Anal. Prev. 138, 105361.
Pantangi, S.S., Ahmed, S.S., Fountas, G., Majka, K., Anastasopoulos, P.C., 2021. Do high visibility crosswalks improve pedestrian safety? A correlated grouped random parameters approach using naturalistic driving study data. Anal. Method. Accid. Res. 30, 100155.
Pearson, K., 1896. Mathematical contributions to the theory of evolution III. Regression, heredity and Panmixia. Philosoph. Transac. Roy. Soc. London Series A 187, 253–318. https://doi.org/10.1098/rsta.1896.0007.
Piccinini, G.F.B., Moretto, C., Zhou, H.P., Itoh, M., 2018. Influence of oncoming traffic on drivers' overtaking of cyclists. Transport. Res. F: Traffic Psychol. Behav. 59, 378–388. https://doi.org/10.1016/j.trf.2018.09.009.
Prati, G., Pietrantoni, L., Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accid. Anal. Prev. 101, 44–54. https://doi.org/10.1016/j.aap.2017.01.008.

Rash-ha Wahi, R., Haworth, N., Debnath, A.K., King, M., 2018. Influence of Type of Traffic Control on Injury Severity in Bicycle-Motor Vehicle Crashes at Intersections. Transp. Res. Rec. 2672 (38), 199–209. https://doi.org/10.1177/0361198118773576.

Rella Riccardi, M., Galante, F., Scarano, A., Montella, A., 2022a. Econometric and machine learning methods to identify pedestrian crash patterns. Sustainability 14, 15471. https://doi.org/10.3390/su142215471.

Rella Riccardi, M., Mauriello, F., Sarkar, S., Galante, F., Scarano, A., Montella, A., 2022b. Parametric and Non-Parametric Analyses for Pedestrian Crash Severity Prediction in Great Britain. Sustainability 14 (6), 3188. https://doi.org/10.3390/su14063188.

Rella Riccardi, M., Mauriello, F., Scarano, A., Montella, A., 2023. Analysis of contributory factors of fatal pedestrian crashes by mixed logit model and association rules. Int. J. Inj. Contr. Saf. Promot. 30 (2), 195–209.

Roberts, E., Chen, T.D., 2017. The effect of crash characteristics on cyclist injuries: An analysis of Virginia automobile-bicycle crash data. Accid. Anal. Prev. 104, 165–173. https://doi.org/10.1016/j.aap.2017.04.020.

The Royal Society for the Prevention of Accidents (RoSPA), 2015. Cycling, RoSPA Policy Paper. Available at https://councilmeetings.lewisham.gov.uk/documents/s35588/05RoSPACyclingPolicyPaper160415.pdf.

Salon, D., McIntyre, A., 2018. Determinants of pedestrian and bicyclist crash severity by party at fault in San Francisco, CA. Accident. Anal. Prevent. 110, 149–160.

Samerei, S.A., Aghabayk, K., Shiwakoti, N., Mohammadi, A., 2021. Using latent class clustering and binary logistic regression to model australian cyclist injury severity in motor vehicle-bicycle crashes. J. Saf. Res. 79, 246–256.

Scarano, A., Aria, M., Mauriello, F., Rella Riccardi, M., Montella, A., 2023. Systematic literature review of 10 years of cyclist safety research. Accid. Anal. Prev. 184 https://doi.org/10.1016/j.aap.2023.106996.

Scholes, S., Wardlaw, M., Anciaes, P., Heydecker, B., Mindell, J.S., 2018. Fatality rates associated with driving and cycling for all road users in Great Britain 2005–2013. J. Transp. Health 8, 321–333. https://doi.org/10.1016/j.jth.2017.11.143.

Scholliers, J., van Sambeek, M., Moerman, K., 2017. Integration of vulnerable road users in cooperative ITS systems. Eur. Transp. Res. Rev. 9, 15. https://doi.org/10.1007/s12544-017-0230-3.

Sivasankaran, S.K., Balasubramanian, V., 2020. Exploring the severity of bicycle-vehicle crashes using latent class clustering approach in India. J. Saf. Res. 72, 127–138.

Sun, Z., Xing, Y., Gu, X., Chen, Y., 2022a. Influence factors on injury severity of bicycle-motor vehicle crashes: A two-stage comparative analysis of urban and suburban areas in Beijing. Traffic Inj. Prev. 23, 118–124. https://doi.org/10.1080/15389588.2021.2024523.

Sun, Z., Xing, Y., Wang, J., Gu, X., Lu, H., Chen, Y., 2022b. Exploring injury severity of bicycle-motor vehicle crashes: A two-stage approach integrating latent class analysis and random parameter logit model. J. Transport. Saf. Security 14 (11), 1838–1864. https://doi.org/10.1080/19439962.2021.1971814.

Thomas, L., Nordback, K., Sanders, R., 2019. Bicyclist Crash Types on National, State, and Local Levels: A New Look. Transp. Res. Rec. 2673 (6), 664–676. https://doi.org/10.1177/0361198119849056.

Townsend, P., 1979. Poverty in the United Kingdom.

Tuckel, P., 2021. Recent trends and demographics of pedestrians injured in collisions with cyclists. J. Saf. Res. 76, 146–153.

United Nations, 2015. Transforming our World: the 2030 Agenda for Sustainable Development. Available at https://sdgs.un.org/.

van Haperen, W., Daniels, S., De Ceunynck, T., Saunier, N., Brijs, T., Wets, G., 2018. Yielding behavior and traffic conflicts at cyclist crossing facilities on channelized right-turn lanes. Transport. Res. Part F-Traffic Psychol. Behav. 55, 272–281. https://doi.org/10.1016/j.trf.2018.03.012.

Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS One 14 (4). https://doi.org/10.1371/journal.

Walter, S.R., Olivier, J., Churches, T., Grzebieta, R., 2013. The impact of compulsory helmet legislation on cyclist head injuries in new south wales, australia: a response. Accid. Anal. Prev. 52, 204–209.

Wang, T., Chen, J., Wang, C., Ye, X.F., 2018. Understand e-bicyclist safety in China: Crash severity modeling using a generalized ordered logit model. Adv. Mechan. Eng., 10 (6), doi: 10.1177/1687814018781625.

Wang, C., Lu, L., Lu, J., 2015. Statistical analysis of bicyclists' injury severity at unsignalized intersections. Traffic Inj. Prev. 16 (5), 507–512.

Wang, Z., Neitzel, R.L., Zheng, W., Wang, D., Xue, X., Jiang, G., 2021. Road safety situation of electric bike riders: a cross-sectional study in courier and take-out food delivery population. Traffic Inj. Prev. 22 (7), 564–569.

Weber, T., Scaramuzza, G., Schmitt, K.-U., 2014. Evaluation of e-bike accidents in switzerland. Accid. Anal. Prev. 73, 47–52.

World Health Organization (WHO), 2020. Cyclist safety, an information resource for decision-makers and practitioners.

Wu, Y., Guo, Yy., Lu, J, 2019. Modeling e-bike crash severity by accounting for unobserved heterogeneity in China. Cictp 2019: Transportation in China-Connecting the World, doi: 10.1061/9780784482292.331.

Wu, C., Yao, L., Zhang, K., 2012. The red-light running behavior of electric bike riders and cyclists at urban intersections in China: an observational study. Accid. Anal. Prev. 49, 186–192. https://doi.org/10.1016/j.aap.2011.06.001.

Xu, B., Huang, J. Z., Williams, G., Wang, Q., Ye, Y., 2012. Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces. Int. J. Data Warehous. Min. 8(2), 44-63. Doi: 10.4018/jdwm.2012040103.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., Wu, C., 2011. Motor vehicle–bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity. Accid. Anal. Prev. 43 (5), 1751–1758. https://doi.org/10.1016/j.aap.2011.04.006.

Ye, F., Cheng, W., Wang, C., Liu, H., Bai, J., 2021a. Investigating the severity of expressway crash based on the random parameter logit model accounting for unobserved heterogeneity. Adv. Mechan. Eng., 13, doi: 10.1177/16878140211067278.

Ye, F., Wang, C., Cheng, W., Liu, H., Vitetta, A., 2021b. Exploring Factors Associated with Cyclist Injury Severity in Vehicle-Electric Bicycle Crashes Based on a Random Parameter Logit Model. J. Adv. Transp. 2021, 1–12.

Yu, C., Xu, M., 2018. Local Variations in the Impacts of Built Environments on Traffic Safety. J. Plan. Educ. Res. 38 (3), 314–318. https://doi.org/10.1177/0739456X17696035.

Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119, 274–289. https://doi.org/10.1016/j.aap.2018.07.031.