Deep Neural Network-based hybrid modelling for development of the Cyclist Infrastructure safety model

Faheem Ahmed Malik[1], Laurent Dala, and Krishna Busawon

Mechanical and Construction Engineering, Northumbria University, Newcastle upon Tyne, NE1 8ST, faheem.malik@northumbria.ac.uk

Abstract. This paper is concerned with modelling cyclist road safety by considering various factors including infrastructure, spatial, personal and environmental variables affecting cycling safety. Age is one of the personal attributes, reported to be a significant critical variable affecting safety. However, very few works in the literature deal with such a problem or undertaking modelling of this variable. In this work, we propose a hybrid approach by combining statistical and supervised deep learning with neural network classifier, and gradient descent backpropagation error function for road safety investigation. The study area of Tyne and Wear County in the north-east of England is used as a case study. An accurate dynamic road safety model is constructed, and an understanding of the key parameters affecting the cyclist safety is developed. It is hoped that this research will help in reducing the cyclist crash and contribute towards sustainable integrated cycling transportation system, by making use of cut above methodologies such as deep learning neural network.

Keywords: Cyclist safety, Deep learning neural network, Safety modelling, Age, Infrastructure.

Declaration: There are no conflicts of interest associated with this publication, and there has been no financial support for this work that could have influenced its outcome.

1 Introduction.

Cycling as a mode of travel has social, economic and environmental benefits. However, it is perceived as a "Risky Activity" [1]. For a cyclist, the interaction between the cyclist and road environment is the essential factor that affects its safety. They face a disproportionate share of risk on roads, e.g., in the UK risk faced by the cyclists in terms of slight crashes per billion vehicle miles is 4,450; highest amongst any road user, and 12.5 times higher than car user for the same traversed distance. Transportation contributes to 25% of greenhouse gas emissions. It is essential to decrease the emissions, which can be achieved through a modal shift towards greener mode of travel, such as cycling.

[1] Corresponding Author : Faheem Ahmed Malik, 209 Wynne Jones Building, Mechanical and Construction Engineering, Northumbria University, NE1 8ST.

For this, it is paramount to increase the safety of cycling as a mode of travel, as its safety and mode share are correlated (see [1, 2]).

Identifying the physical and environmental threats to cyclist safety within the network allows a critical insight into the cyclists' preference and choice [3]. The built environment, weather, work-related factors, and attitudes affect the everyday commute by bicycle [4]. Cycling hazards are also dependent upon cyclist-specific variables of age, experience, and gender [1]. A route choice study carried out in Texas [5], concluded that the cyclist route choice depends on the attributes of the route and cyclist's demographics [5]. The cyclist's route network preference varies with its personal attributes and behaviour of other road users [6]. The work on cyclist near misses in London [7] led them to conclude that the rider's age group directly affects their daily near misses. The number of incidents per day decreases with age, from 2.47 (20-29) to 1.85 (> 60 groups). These near misses are found to be correlated with the crashes (see [8, 9]). The similar results were obtained in Germany, wherein it was concluded that the cyclists of different age group use the infrastructure differently, and exhibit different microscopic road traffic behaviour [10]. The study in Palermo city (Italy) to investigate associations between severity of non-fatal crashes and driver characteristic reported that riders below 25 years are more likely to be involved in a slight or serious crash than riders from any other age group, followed by elderly population (greater than 64) [11]. The similar study in Sweden to investigate the cyclists' injury by age and gender found that the elderly population is at a relatively higher risk than the middle-age population, with a much more significant fatal risk for elderly women [12]. An analysis of modal shift scenarios of short tips to cycling and effect on overall road safety in Netherland led them to conclude that mode shift can substantially affect road safety for different age groups. The most notable impact was modelled on the elderly population, for which the risk is expected to increase significantly [13]. However, presently there is insufficient evidence to understand the relationship between cyclist safety and these identified variables [14], due to the modelling inability. Cycling safety is an important topic, but there are limited studies which explore the cycling risk to their exposure [15]. Additionally, there is a need for the capabilities to assess the safety of the experimental roadway designs and/or operational strategies before they are built or employed in the field [16]. Therefore, the present research needs to develop a road safety model that considers this dynamic variation of safety. Such a model should model the safety based upon the rider's attribute and should be operable even in the initial planning and design of the cycling network.

To model safety, the first mathematical theory to be used is generalized linear modelling. Over time, various studies proposed a generalized linear model with the assumption of a non-normal error structure [17]. This overcame the limitations associated with the linear regression models and produced a better fit to the observed collision data [18]. As the crashes are discrete positive integral variables, therefore this prompted the use of Poisson regression. However, it is unable to handle overdispersion (i.e. the variance exceeding the mean). This then motivated using negative binomial or Poisson gamma models, assuming that the Poisson parameters follow a gamma distribution [19]. However, there are locations with zero reported crashes, this motivated the use of zero-inflated Poisson method, having two different states; zero state and normal count state. For improving the modelling capabilities, various techniques such as hierarchical, random effect, cart, finite-mixture/latent-class, log-linear, probit/logit, Markov switching, Poisson–Log normal Regression, Empirical Bayes Method, Conway-Maxwell-Poisson, negative binomial-Lindley method and others [20, 21], have been explored in the literature. However, all the present available crash models are reactive and cannot consider the dynamic nature of the cyclist's interaction with variable infrastructure and quantify its safety implications. These all are based upon modelling the human error, without considering the cyclist's vulnerability, and its susceptibility to various externalities.

This paper aims to develop a fundamental understanding of one of the reported dynamic variables: the trip maker's personal attribute, i.e., the rider's age. This is motivated by the fact that this variable has been reported as a significant variable in the literature. Still, there are very few works which deal with modelling this variable. Besides, it is shown that motorists exhibit behavioural sensitivity to the bicyclist appearance [22]. Consequently, we seek to understand how the rider's age affects their safety in the natural road environment. By modelling this variable, it is expected that the knowledge obtained can be utilized for better design and planning of cycling infrastructure based upon its intended users. We propose a knowledge-driven approach for infrastructure planning based upon the specific users rather than the infrastructure's generalized usage. More precisely, our objectives are:

- To develop an understanding of how safety is affected by the age group of the rider.
- Test the hypothesis that unsafeness of the interaction between user and the infrastructure depends on the user's age.
- To develop a dynamic safety model with age as an output variable.
- Identify the most important variables affecting the unsafeness of an age group.
- Validate the importance of the identified variables statistically.

In the next section, we describe the considered study area for the proposed research. In Section 3, the proposed methodology is described. In Section 4, the results of the research are presented, followed by discussion. Finally, some conclusions are drawn in Section 5.

2 The considered area of study

To achieve the aim and objectives, northeast of England (Tyne and Wear County) is selected as the considered area of study. It is one of the nine official regions of England, encompassing an area of 3,317 sq. miles, population of 1.13 million, housing five boroughs Gateshead, Newcastle–upon-Tyne, North Tyneside, South Tyneside and Sunderland, thirteen urban and three rural districts.

The Department for Transport (DfT) houses the database for road crashes in the United Kingdom. For each road traffic collision, a trained road crash investigator visits the crash site and records the crash in a document known as STATS 19, consisting of four sections: i) Accident Statistics, ii) Vehicle Record, iii) Casualty Record, and iv) Contributory Factors. The Gateshead city council provided access to the crash database. The accessed dataset houses: i) Type of severity, ii) Time, date and location of the crash, iii) Environment conditions such as lighting conditions, weather, road surface conditi-

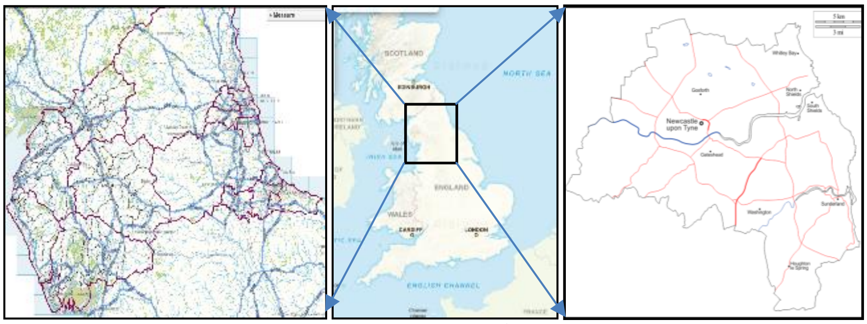


Fig. a Location and Boundaries of the study area

-on, type of infrastructure and number of vehicles involved, iv) Sociodemographic information such as age and gender of the cyclist. The classification of the severity is performed through the Department for Transport (DfT) criterion. A crash is classified fatal; if the crash results in the death within 28 days from the crash, serious; if it results in either a death after 28 days or at least an overnight admission in the hospital and slight; if crash results in overnight discharge from the hospital or property damage only [23]. The crash investigation by DfT aims to record the information as accurately as possible, as it serves the basis for further legal and other courses of actions

There are 3,325 bicyclist crashes recorded in the study area between 2005 and 2018. Out of these, 79.3% are slight, 19.9% serious and 0.8% are fatal crashes. The subsequent crash distribution is obtained for the respective age groups.

Table 1. Age distribution of the crashes

Age	**Frequency**	**Per cent**	**Age**	**Frequency**	**Per cent**
Under 17	1420	42.7	45-54	251	7.5
17-24	537	16.2	55-64	115	3.5
25-34	494	14.9	Over 64	65	2.0
35-44	347	10.4	Unknown	96	2.8

3 Hybrid modelling for cycling safety.

Besides developing a predictive safety model, the investigation will uncover the causality, have a high predictive capability and be scalable to a large data set. Therefore, a combination of the data-driven and statistical methods seems to be the most appropriate for such an investigation. This work proposes a hybrid model combining: a) Traditional Safety, b) Causal Inference, and c) Data-driven methods. Firstly, the traditional statistical models are constructed, using crash and mode share risk rate. Then for causal inference, heat maps are developed to understand the inference between infrastructure and the age groups. Deep learning is used to construct the riskiest age prediction model and identify the variable importance of the different input variables. To validate the inference obtained, results are validated using the chi-square test for testing the association between age and identified variables. The strength of the inference is tested using Cramer's V value. This framework consists of the following methodologies:

3.1 Statistical Risk Rate and heat maps

The mode share (miles share rate) of each age group is calculated by scrutinizing and evaluating Department for Transport's National Travel Survey (NTS) database. This is a household survey collected through household interviews and trip diaries, the primary source of data on England's personal travel pattern. A base input file for crashes is constructed, having detailed information regarding each crash. In the next step, crashes are grouped based upon the age of the rider involved and evaluated. Then both crash and mile rates are compared and gauged, to calculate the risk faced by each group (for the same distance traversed). To compare the age groups within themselves, the normalized risk is determined for each age group, concerning the safest age group. The analysis is performed accurately up to one decimal place. For investigating the spatial variation of risk with different infrastructure, heat maps are generated for the identified age groups.

3.2 Deep Learning Neural network.

A predictive model is developed by using deep learning with neural network classifier, and gradient descent backpropagation error function. It is the sub-group of a machine learning techniques based upon computational methodologies which imitate the working of the human brain. The neural networks were introduced firstly in transportation research in the 1990s [24]. The road safety problem is highly non-linear and characterized by the underlying correlation between various infrastructural, environmental, and personal attributes of the rider. The neural network has been widely applied as an analytical data method in this field [25], as these result in generic, accurate, and convenient mathematical models, which can simulate the numerical model components [26]. This is due to their ability to work with a large amount of multi-dimensional data, modelling flexibility, learning, generalization ability, adaptability and good predictive capacity [26]. The primary motivation for employing deep learning for safety modelling is that crashes are highly non-linear. The modeller has no guidance from either theory or even dimensional analysis for modelling. Although other algorithms exist and deep learning neural networks are not a new concept, its ability to solve the complex and the interchangeable system problems, which the transportation system is characterized by, is the rationale for its use.

In the first step of building neural model, a learning algorithm is developed to divide the data set randomly into training (65%), validation (30%), and testing (5%). This division ensures proper learning of the constructed model, assesses the trained model and ensures that the constructed model is relevant to untrained scenarios [25, 27]. The predictive safety model is developed using four input variable types: a) Infrastructure, b) Spatial, c) Personal, and d) Environmental input variables (Table 2).

Table 2. Input variable for the proposed model.

No.	Input Variable	Values
1.	**Infrastructure**	
a).	Speed limit (Maximum permissible speed limit on the road).	20-70

b).	1st Road Class (For intersections, the rider may be required to move from one hierarchy level of road classification to another. This is the first hierarchy classification of the road from which the rider is moving towards the next one).	A, B, C, E, U
c).	2nd Road Class (Hierarchy classification of the road that the rider to intending to move to / already moved to).	A, B, C, E, U
d).	Junction Detail (Type of intersection).	Crossroad, Mini Roundabout, Multiple Junction, Straight Road, Roundabout, Slip Road, T or Staggered, Private Drive
e).	Junction Control (Type of control employed at the intersection).	No Control, Traffic Signal, Give way or uncontrolled, Stop sign
f)	Vehicle Maneuver (Maneuver that rider was performing/intending to perform when the crash occurred).	Changing lanes, Going ahead, Moving off, Overtaking, Parked, Reversing, Slowing/stopping, Turning, U-turn, Waiting to go ahead, waiting to turn
g)	Carriageway Hazard (Additional unexpected hazards on the carriageway).	Animal in the carriageway, Dislodged vehicle load on the carriageway, None, Object in the carriageway, Pedestrian on the carriageway.
h)	Road Type (Type of road infrastructure present at crash spot).	Dual Carriageway, One-way street, Roundabout, single carriageway, slip road,
i)	Vehicle Junction Location (Location of cyclist at the junction, when crash occurred).	Approaching junction or waiting/parked at junction exit, cleared junction or waiting/parked at junction exit, Entering, Leaving, Mid Junction, Straight Road (Not at or within 20 meters of the junction)
j)	Road Location of vehicle (Location of cyclist to the road infrastructure, when crash has occurred).	Bus Lane, Busway, Cycle lane, cycleway, footpath, on layby or hard shoulder, main carriageway, tram/light rail track
k)	Skidding and Overturning (After crash whether there was any skidding or overturning).	No skidding or overturning or jack-knifing, overturned, skidded, overturned, and skidded
l)	Special Conditions at site (Any infrastructure defects at crash location).	Defective Traffic Signal, None, Oil, mud, defective road signs or marking, defective road surface, roadworks,
2.	**Spatial**	
a).	Journey Hour (The hour in which crash occurred)	0-23
b).	Number of vehicles (Number of vehicles involved in the crash).	1-5
c).	Month of Journey (Month in which crash occurred).	Jan-Dec
d)	Journey Day (Day of week on which crash occurred. The day, month and hour of journey are a representation of the traffic flow regime plying at the time of the crash)	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
3.	**Personal attributes**	
a).	Gender (Gender of the rider).	Male, Female and Unknown
b).	Breath Test (To check whether rider was intoxicated or not).	Negative, Positive and Not Required
c).	Journey Purpose (The purpose of journey being undertaken).	Commuting, work trip, School Journey by Pupil, taking pupil to school, other, Unknown
4.	**Environmental**	
a).	Lighting conditions (The lighting conditions, and presence and working of streetlights).	Daylight /Darkness- No Street Lighting, Street Lighting Unknown, Street Lights present and lit, Street Lights present but unlit,
b).	Meteorological conditions (The meteorological conditions when the crash occurred).	Fine/Rain/Snow-with high winds, without high winds, fog or Mist Hazard, Other.
c).	Road Surface Condition (The road surface condition at the time of the crash. The road surface and meteorological conditions may not necessarily be the same).	Dry, Frost/ice, Wet/damp, Snow
	Output Variable	**Riskiest Age Group (0-17, 14-24,25-34, 35-44, 45-54,55-64, and over 65).**

Considering that relationship between input variables and output is highly non-linear and complex [28]; therefore, two hidden layers are used in the network. The batch training, cross-entropy error function, and scaled conjugate gradient optimisation are used. The network structure is explicitly defined in Table 3.

Table 3. The network structure of the deep learning model.

Network Topology	Number of hidden layers	2
	Elements in each layer	350
	Activation function between the hidden layers	Hyperbolic Tangent
	Activation function between hidden and output layer	SoftMax
Training	Type of Learning	Supervised
	Optimization	Gradient Descent (Batch)
	Iterative Method	Scaled conjugate gradient
	Initial Lambda	0.000000001
	Initial Sigma	0.000000001
	Initial Centre	0
	Initial offset	±0.000000001
Stopping and Memory Criterion	Steps (maximum) without a change in the error	999,999
	Training (maximum) time	999,999
	Training (maximum) epochs	999,999
	Relative change in the training error (minimum)	0.000001
	Relative change in the training error ratio (minimum)	0.000001
	Cases to store in the memory (maximum)	999,999

The following four-step iterative approach is used for modelling each of the input variables with the output.

Step 1: The random weights are assigned to each of the input variable connection (between the input and hidden, first and second hidden, and between the hidden and output layer).

The activation function' Hyperbolic tangent' is used for developing the weights in hidden layers, given by:

$$O_j = \tanh\left(S_j\right) = \frac{e^{S_j} - e^{-S_j}}{e^{S_j} + e^{-S_j}} \tag{1}$$

In the output layer, activation function 'SoftMax' is used, given by:

$$O_j = \sigma\left(S_j\right) = \frac{e^{S_j}}{\sum_{k=1}^{m} e^{S_k}} \tag{2}$$

m is the number of output neurons, and O_j is the activation of the jth neuron

These functions take real numbers as arguments and return real values [-1, +1].

Step 2: The error between the desired output (target) and output obtained, is calculated using cross-entropy error function, given by:

$$E = -\sum_{j=1}^{m} t_j \ln O_a \tag{3}$$

O_a is the actual output value of the output node j, t_j is the largest value j,and m is the number of output nodes.

Step 3: Based on the error (step 2), the initial synaptic weights are updated. In each epoch, the backpropagation algorithm calculates the gradient of the training error as:

i) nodes between input and hidden layer:

$$\frac{\partial E}{\partial w_{hj}} = \sum_{j=1}^{m} \left(O_a - t_j\right) x_h w_{hj} (1 - x_h) x_i \tag{4}$$

ii) nodes between output and hidden layer:

$$\frac{\partial E}{\partial w_{hj}} = (O_a - t_j)x_h \tag{5}$$

In each of the training case (epoch), the weight w_{ih} is updated by adding it:

$$\Delta\, w_{ih} = -\gamma \frac{\partial E}{\partial w_{hj}} \tag{6}$$

$$\Delta\, w_{ih+1} = w_{ih} + \Delta\, w_{ih} \tag{7}$$

x is the input variable, γ is the learning rate, and w_{hj} is the synaptic weight for the jth neuron.

Step 4: Iteration (scaled conjugate gradient): The updating of weights is iterated until either the minimum change in the training error or the maximum number of these iterations (epochs) is achieved.

The recommended methodology to measure the neural models' performance is through Receiver Operating Characteristics (ROC) curve [25], which gives the visual display of sensitivity and specificity for all the possible cut-offs. The Area Under the Curve of the Receiver Operating Characteristics (AUROC) quantifies the model's performance, resulting in an evaluation matrix used to evaluate networks' classification performance. ROC is a probability curve, and AUROCC represents the measure of the separability power of the network. Higher the AUROC value, the network's distinguishable power between the risky and non-risky age groups is better. Besides, gain and lift charts are used for qualitative evaluation, the visual aids for evaluating the performance. The model is then validated through validation datasets, ensuring an unbiased review of the model fit on the validation dataset while tuning model hyperparameters. Thereupon model's performance is checked on unseen data, providing an impartial evaluation of the final model constructed based upon the testing dataset. Through this three-step process of training, validation, and testing, the constructed model's performance is estimated to establish the credibility and confidence for further evaluation, planning, design, and policy implications.

The critical variables in the data learning model are identified through variable importance. Each variable's normalized importance concerning the most critical variable is also evaluated to compare variables relative to each other. This is based upon both testing and validation data sets. The independent variable importance is a measure of how much the predicted output value changes, viz a viz change in the input variable. Each input variable's normalised importance is their respective importance value divided by the largest importance value and expressed as percentages.

3.3 Chi-Square Test

After developing the predictive model, the statistical validation of the identified critical variables is undertaken. The input variables affecting the crashes are measured either on a nominal or ordinal scale. Therefore, the non-parametric technique is the ideal statistical method in such scenarios, especially when the sample size is small. The two assumptions of: i) Samples being random, and ii) Observations being independent of each other [29] need to be met. The crashes are a random phenomenon [30] and are independent of other crashes occurring at different locations, thereby, satisfying the two pre-requisites. Chi-square test for goodness of fit, a non-parametric technique, specifically designed to solve such complex non-linear problems, tests whether there exists a relationship between two variables and uses the sample data to test the hypothesis regarding the shape of the proportion of population distribution. It determines how well obtained sample proportions fit the population proportion specified by the null hypothesis. Each variable in the sample is classified on n variables, creating an n-dimensional frequency distribution matrix. As the matrix is greater than two by two order, a modification of the Phi-Coefficient, known as Cramer's V, is used to measure the strength of association [31]. The following four-step procedure is used.

Step 1: Chi-square statistic is calculated, as:

$$\chi^2 = \sum \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}} = \sum \frac{(Observed - \ Expected)^2}{Expected} \tag{8}$$

Step 2: Degree of freedom of the two variables, whose association being evaluated is calculated, as:

$$df = smaller\ of\ either\ (R-1)\ or\ (C-1) \tag{9}$$
$$where\ R\ is\ number\ of\ rows, and\ C\ is\ the\ number\ of\ columns.$$

Step 3: For determining the strength of the correlation, Cramer's V statistic is used, a post-test (after Chi-square correlation test):

$$V = \sqrt{\frac{\chi^2}{n(df)}} \tag{10}$$

Step 4: Cramer's V is a single-valued numeric output, which needs to be converted into qualitative knowledge, performed using Cohen's table. This determines the strength of correlation using the degree of freedom and the numerical V value, in terms of no correlation, small, medium and large correlation.

4 Result and Discussion

4.1 Risk Rate and Heat Maps

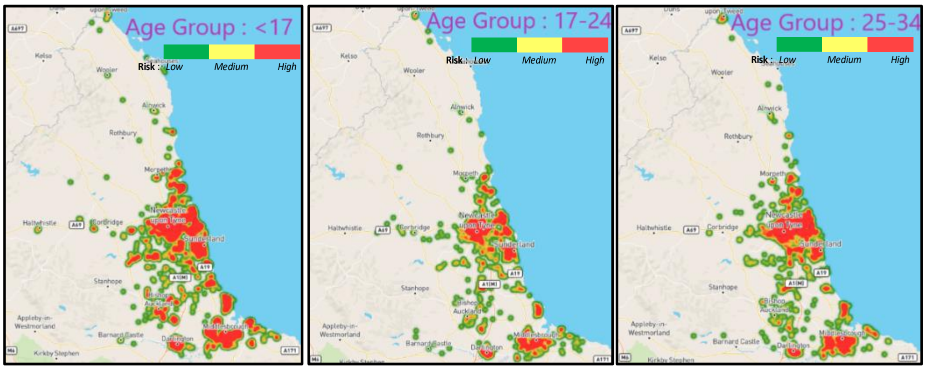
The riders age group are divided into eight groups (0-16 to >70), the relative risk is calculated for each group based upon the crash rate, and their relative distance travelled. The corresponding normalized risk is calculated, with respect to the safest age group age (60-69).

Table 4. Age distribution of the crashes

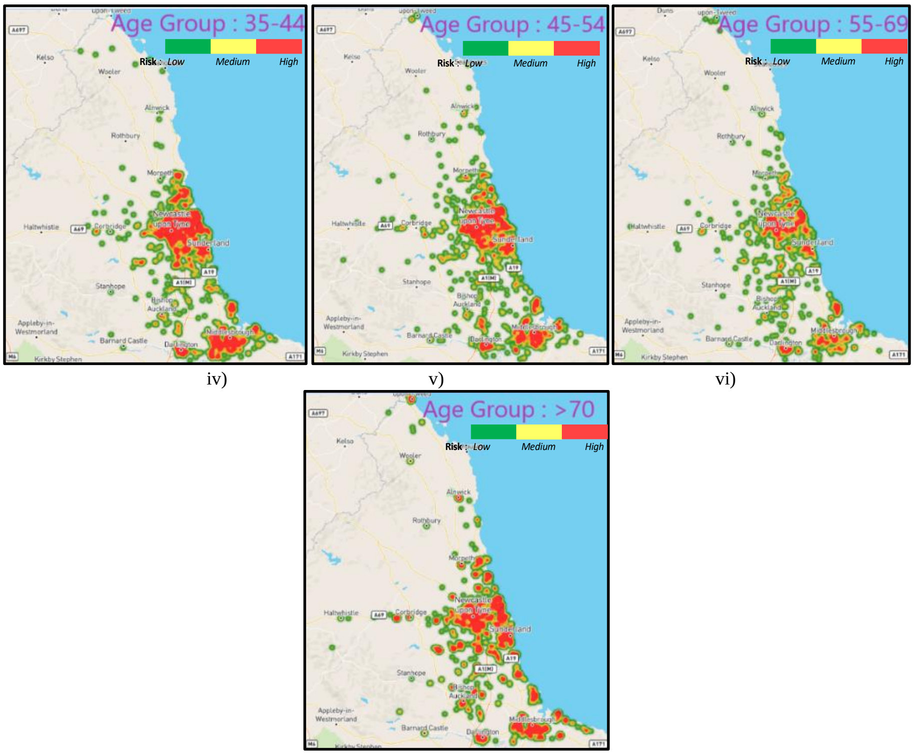
Age	Miles share rate	Crash rate	Relative Risk	Normalized Risk
0-16	5.9	43.0	7.3	26.9
17-20	9.2	9.8	1.1	3.9
21-29	13.1	14.5	1.1	4.1

30-39	15.1	13.1	0.9	3.2
40-49	23.0	9.8	0.4	1.6
50-59	19.0	5.4	0.3	1.1
60-69	9.7	2.6	0.3	1.0
70+	5.0	1.8	0.4	1.3
Total	100	100		

The risk rates and normalized risk lead to infer that the cyclist's risk decreases with age. The risk faced by the age group under 17 is 27 times higher than the age group of 60-69 for the same distance traversed. The risk for the cyclist continues to decrease with age, from 17 to 69. However, the elderly population (age >70) face a proportionally higher risk than the two preceding age groups. This can be attributed to physical and cognition limitation with advanced age. These results agree with the results obtained in other European countries. Similar results for the young and elderly population were obtained in Italy [11] and Netherlands [32]. In the UK, London's naturalistic study found everyday near-miss incidence rate for cyclist's decreases with the rider's age [7]. We can thus conclude that the risk for the cyclists decreases with the age of the rider. There are underlying factors which contribute to a decrease in normalized risk with age. These include a reduction in risk-taking behaviour with age, better control, experience, and behavioural sensitivities of other road users with the rider's appearance. The motorists have been found to exhibit behavioural sensitivity to the bicyclist appearance [22] and change their behaviour of interaction with the cyclist based upon the riders' own attributes. Therefore, age is a multilayer variable affecting the safety of cyclist in multiple ways. To test the hypothesis, that unsafeness of the interaction between the rider and infrastructure depends on the age of the user, following risk heat maps are generated for each age group in the investigation area.



i) ii) iii)



iv) v) vi)

vii)

Fig. b Hotspot identification: i) under 17, ii) 17-24, iii) 25-34, iv) 35-44, v) 45-54, vi) 55-69, vii) over 70

The heat maps demonstrate that the risk that infrastructure present to riders is dependent upon their age. There is an expected centralization in Newcastle city centre, as it has a higher cyclist flow than other parts of the study area. The similar results for the city centre, have been reported in the literature for university towns (see [33]). For the rest of the study area, the pattern and the spread of the crashes, are different for different age groups. The naturalistic study on cyclists in Germany found that microscopic traffic parameters are significantly different for riders belonging to different age groups [10]. There are location-specific infrastructure parameters which determine the risk, affecting cyclists differently. The cyclist's attributes also influence their interaction with the infrastructure, i.e. the same infrastructure can pose a different risk to different riders. Therefore, we can conclude that not only infrastructure is a dynamic variable, but also the age of the rider is also a dynamic variable affecting its safety.

The findings are contrary to the variables modelled in the present road safety models. The critical variables modelled in American/Canadian crash prediction model is the Annual Average Daily Traffic (AADT) on minor and major road [34]**.** British crash prediction model, takes AADT and the length of the investigated infrastructure, as input variables [35]. Similarly, Danish model takes AADT and road geometry [36]. Land use pattern and hierarchy of road are the variables considered by the Swedish crash prediction model [37]. TRAVA, i.e., Finnish crash model considers speed limit, number of intersections, lighted, paved road, sight distance, congestion, number of vehicles and percentage of heavy vehicles [38]. These conventional road safety models are ill-equipped to the specific and peculiar needs of the cyclist. An in-depth safety model is developed in the next section for the cyclist, modelling dynamic input variable of 'age of the rider'.

4.2 Deep Learning Results

The constructed deep learning model based upon the identified input critical variables from literature has the following characteristics (Table 5). The output is the riskiest age group.

Table 5. Model features of the constructed deep learning model.

		Sample Size	Per cent
Sample	Training	2142	64.5%
	Validation	903	30.0%
	Holdout	180	5.5%
Total		3225	100.0%
Dependent Variable: Driver Age Group			
Input Layer	Number of Units	173	
Hidden Layer(s)	Number of Hidden Layers	2	
	Number of Units in each Hidden Layer	350	
	Activation Function	Hyperbolic tangent	
	Error Function	Cross-entropy	
	Cross-Entropy Error	2674.1	
Output Layer	Dependent Variables	Driver Age Group	
	Number of Units	7	
	Activation Function	SoftMax	
	Error Function	Cross-entropy	
	Cross-Entropy Error	1162.9	

The ROC curve, gain and lift charts developed for the constructed model are shown in Fig c. The AUROC values presented in Table 6 to establish the credibility of the model.

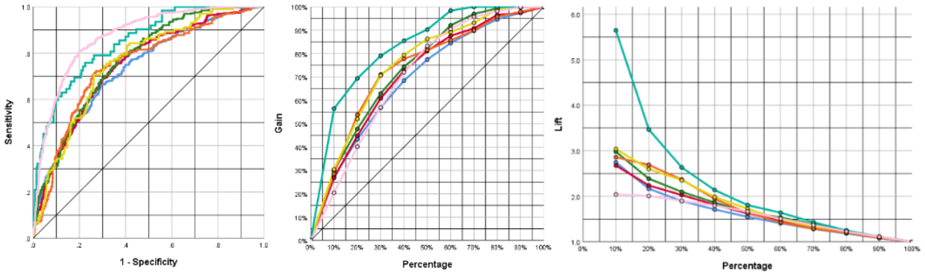


Fig. c: i) ROC curve, ii) Gain chart, and ii) Lift chart for the constructed deep learning model.

Table 6. Area Under Receiver Operating Curve (AUROC) for the output variable

Variable	AUROC	Variable	AUROC

Under 17	0.87	**45-54**	0.75
17-24	0.74	**55-64**	0.76
25-34	0.75	**Over 65**	0.85
35-44	0.77	**Average**	0.81

The AUROCC values obtained for over 65(85%), 55-64(76%), 45-54(75%), 35-44(77%), 25-34 (75%), 17-24 (75%), and under 17 (87%) age groups, indicate a high distinguishable capability between the risky and non-risky age groups. The accuracy achieved is plausible, considering the multifactor nature of crashes. To evaluate the model's prediction capability, gain and lift charts are developed, indicating the model has an excellent prediction capability. Therefore, we can conclude that developed model can be used efficiently for predicting the riskiest age group based upon the specific input variables. There are very few works in literature, which have been able to model the age variable for safety analysis with such reasonable accuracy and efficiency.

(Hossaon and Muromachi, 2009)[39] found that majority of motorist crash prediction models have the prediction success of less than 50%. (Peltola and Kulmala, 2010) [38] found an error of 65% in the Finnish crash prediction mode TRAVA for the cyclist. Similarly, Federal Highway Administration FHWA [16] (transportation department USA) analysis on the safety analysis using the major simulation software's, VISSIM, AIMSUM, TEXAS and PARAMICS revealed that there are modelling inaccuracy in the microsimulations for the cyclist. (Lawson *et al.*, 2013) [40]argued that the conventional models are developed for the assignment of the motorized modes of travel and are not equipped for the cyclist's needs, as these are unable to quantify the effect of the cyclist safety performance function. A survey on safety models [41] found that around 70% of the European road agencies rarely or never systematically use the collision prediction model in their decision making owing to these reasons. The constructed model has superiority over the available traditional road safety models in the literature. This is attributed to the deep learning neural network's ability to model the non-linear and complex relationship between input and output variables.

These present models are usually probability-based. The gain and lift charts evaluate developed model's distinguishable capability compared to a non-model probabilistic approach (baseline scenario). In the gain chart, all the predicted outcomes are higher than the baseline scenario of 45^0, reinforcing the constructed model's appropriateness. The same is depicted in the lift chart, e.g. in predicting the age group > 70 years, at 10% data points, the accuracy of the model is 5.5 times higher than the base case. The developed safety performance functions are equipped to the particular needs of the cyclist. The model does not require historical crash data for modelling. The various input variables of infrastructure, spatial, personal and environmental variables can be directly used to model safety, once the model has been constructed. It can be applied to an infrastructure which is still in the planning and design phase.

The importance of each of the variable and the normalized importance with respect to the most critical variable, are calculated and tabulated in Table 7.

Table 7. Normalized importance of the input variables.

Independent Variable Importance					
Variable	Importance	Normalized Importance	Variable	Importance	Normalized Importance
Journey Purpose of Driver/Rider	0.113	100.0%	Junction Detail	0.039	34.2%
Number of Vehicles	0.066	58.5%	Skidding and Overturning	0.038	33.6%
Hour	0.065	57.4%	Breath Test	0.038	33.3%
Vehicle Manoeuvre	0.059	51.7%	Weather	0.033	29.5%
Carriageway Hazards	0.049	43.6%	Special Conditions at Site	0.03	26.9%
Road Location of Vehicle	0.048	42.0%	Road Surface Condition	0.029	25.3%
Light Conditions	0.045	39.6%	2nd Road Class	0.029	25.2%
Road Type	0.043	38.2%	Day	0.028	24.5%
1st Road Class	0.042	37.1%	Casualty Gender	0.024	21.4%
Speed Limit	0.041	36.3%	Junction Control	0.024	21.4%
Junction Location of Vehicle	0.04	35.7%	Driver Gender	0.022	19.2%
Month	0.039	34.6%	Weekday or Weekend	0.014	12.8%

The most significant variable affecting the risk for an age group is the *rider's journey purpose*. This is followed by the number of vehicles involved in the crash and the hour in which the journey is undertaken; both are spatial variables and represent the *traffic flow regime*. They are followed by vehicle manoeuvre, carriageway hazards, and road location of the cycle, which are infrastructure variables that define cyclist interaction *with the infrastructure*. The *lighting conditions* that the cyclist is subjected to impact cyclists' safety, which varies with the rider's age. This is an expected result as to how different age groups react to different lighting conditions is dependent upon their experience, physical and cognitive capabilities. This is followed by road type and class, and the speed limit of the *infrastructure*. This implies that the riders from different age groups interact differently with the different road infrastructure. The variable importance from the constructed deep learning model and the risk rates and hotspot heat maps led us to conclude that infrastructure poses a different risk to the cyclist based upon the rider's age. The study results can have significant implications on the policy, design, and planning of the road network. The present models do not consider the variable age and are based upon the assumption that road safety is independent of age. The cyclist age distribution is highly varied and can vary significantly from one place to another. Therefore, the research can help develop focused remedial measures to improve safety based on the intended users, rather than the infrastructure's average usage.

4.3 Statistical Modelling

The variables, having the importance >0.04, are selected for further analysis. The association between the target variable and input variables are tested statistically, and their association with safety is determined using Cramer's V value and Cohen's table.

Table 8. Chi-square test, and the statistical association between the significant variables and target variable.

Null Hypothesis H_0	**Alternate Hypothesis H_1**	**Degree of Freedom**	**Pearson Chi-Square** χ^2	**p-value**	**Hypothesis Adopted**	**Type of association**		
Driver Age risk is Independent of	Driver Age risk is dependent on	*df =(R-1) ∧ (C-1)*				Degree of Freedom	Cramer's V	Type of Association
Journey Purpose.		*(7-1) ∧ (6-1) = 5*	520.95	0.01	H_1	5	0.18	Medium

Number of Vehicles.	*(7-1) ∧ (5-1) = 4*	238.69	0.01	H_1	4	0.136	Small
Hour of Journey	*(7-1) ∧ (23-1) = 6*	678.61	0.01	H_1	6	0.187	Medium
Vehicle Manoeuvre.	*(7-1) ∧ (18-1) = 6*	309.68	0.01	H_1	6	0.127	Medium
Carriageway Hazards.	*(7-1) ∧ (5-1) = 4*	75.71	0.01	H_1	4	0.153	Medium
Road Location of Vehicle.	*(7-1) ∧ (8-1) = 7*	190.19	0.01	H_1	7	0.099	Small
Light Conditions	*(7-1) ∧ (7-1) = 6*	203.68	0.01	H_1	6	0.103	Small
Road Type.	*(7-1) ∧ (6-1) = 5*	168.96	0.01	H_1	5	0.103	Small
1st Road Class	*(7-1) ∧ (5-1) = 4*	368.41	0.01	H_1	4	0.169	Medium
Speed Limit.	*(7-1) ∧ (6-1) = 5*	265.44	0.01	H_1	5	0.128	Small

A significant correlation exists between all the identified variables and age group at a 99.9% confidence interval. A medium strength of the correlation is obtained for journey purpose, hour of journey, vehicle manoeuvre, carriageway hazard, and 1st road class. A small strength correlation is obtained for the number of vehicles, road location of vehicle, lighting condition, road type and speed limit. The results indicate no single variable has a high strength of correlation with the age of the rider, which affects its safety. A single high correlation would have been contrary to the established road traffic crash modelling theories [42, 43]. The statistical analysis of the identified variables has validated the results obtained by deep learning neural networks.

4.4 Model Significance

At present, the safety analysis is mainly performed at the macro level, such as country level, and demographics of the intended users are ignored, e.g. a university town such as Oxford, may have a different population demographics than an old English mining town such as Sunderland. The study results demonstrate that if we undertake such modelling without considering the age distribution, it will lead to inaccurate modelling. Hence, a single countrywide model without considering age distribution of the particular area such as a city or a county will lead to improper modelling, and corresponding inaccurate recommendation measures. Such a model may be appropriate for motorists, who benefit from a machine at their disposal. A motorist's physical and cognitive abilities do not get severely strained as a cyclist, nor is the maturity and ability to respond to the riskiest situation a critical safety variable.

Numerous studies have questioned the present modelling and their ability to model the cyclists' idiosyncratic needs [3, 44]. The hybrid methodology proposed and applied in the Tyne and Wear not only models’ the safety accurately, but also develops the understanding of the interaction of the variables, and how they affect safety. These attributes, such as the journey purpose, traffic flow regime, and infrastructure parameters, are all dynamic variables unique to a cyclist. Therefore, there is a need to develop the models specifically for the cyclist using such an intelligent hybrid methodology based upon deep neural networks; demonstrated as an effective method of modelling safety and understanding variable interactions to affect the cyclists' safety. Hence, we can conclude that the present methodologies, such as probability or regression-based, need to be replaced. Such a shift in modelling will result in a better understanding of cycling safety, identifying the crash causation, knowledge-driven recommendation measures, and an integrated sustainable transportation system. Such studies have a renewed focus as we move towards the pathway for the autonomous transportation system. The cyclist's variabilities modelled can be inputted into the V-V (vehicle to vehicle) and V-I (vehicle to infrastructure) algorithm for autonomous vehicles. These algorithms will consider the rider's variability in a specific age group at the critical infrastructure type or the particular environmental/ spatial conditions.

The local authorities can also use the model to plan, design, and optimize the cycling network based upon the intended population (age distribution) and model the safety considering the infrastructure, environmental, spatial and other personal attributes of gender and journey purpose. Therefore, this model also considers the land use pattern, the peak, staggered peak, and other dynamic variables varying from a city to city. Even, through inverse analysis based upon the rider's age, the model will predict the riskiest infrastructure variables keeping the environmental, spatial, and personal attributes constant for a particular scenario. This can be performed for different age groups, and then combined using the optimization algorithms (scaled conjugate gradient) to predict the riskiest and safest infrastructure type.

The model can be interoperable to a different city/ country, as cycling safety factors are not expected to change significantly. However, there may be variation in the significance importance of the variables. Therefore, before applying the model to different scenarios, it needs to be validated, similar to all the major simulation packages.

5 Conclusion

A cyclist is a vulnerable road user. The manner of interaction of cyclist with the road infrastructure depends on several factors, including cyclist's own personal attribute, i.e., the rider's age. The present crash models are mostly developed for motorists in general, without considering cyclists' limitations. A dynamic hybrid approach is applied in this research. The causal relationship between the variables and riders' age is identified and statistically validated without compromising the accuracy or predictive capability.

The study has demonstrated the superiority of the supervised deep learning neural network, over other traditional mathematical theories by modelling the dynamic variable, i.e., the rider's age effectively and efficiently. An accurate dynamic road safety model has been constructed, and an understanding of the key parameters affecting the cyclist safety has been developed. The following main conclusions are drawn from this study:

- The cyclist's risk decreases with age, e.g. riders under the age of 17 are 27 times more likely to be involved in a crash than the age group of 60-69 for the same distance traversed.
- There is no single variable having a high strength of correlation for road safety with the rider's age, reinforcing that cycling safety is a multifactor and multidimensional phenomenon, requiring a similar modelling approach.
- Different infrastructure network pose a risk differently to riders belonging to different age groups.
- The age of the rider influences other road user's interaction with the cyclist.

- The unsafeness of the interaction between the rider and infrastructure is dependent upon the age of the rider. This interaction is dependent upon a variety of dynamic variables in the following descending order:
  (a) Personal Characteristics (Journey Purpose),
  (b) Traffic flow regime (Number of vehicles, and hour of travel),
  (c) Manner of Interaction of the cyclist with the infrastructure (vehicle manoeuvre, carriageway hazards and road location of the vehicle),
  (d) Environmental (lighting) conditions,
  (e) Infrastructure variables (road type and class, and speed limit).

The present research in the road safety modelling needs to move from the simple probability-based models to deep learning neural models, which can open up new possibilities, as demonstrated in this work. The study results can significantly impact the route choice, modelling, and planning of infrastructure. The constructed model can assess with certainty regarding the type of infrastructure required to increase safety, based upon the indented users rather than a generalized approach. This can be even employed to an infrastructure which is still in its planning/design phase, considers the vulnerability of rider, its susceptibility to externalities, and the varied safety effect based upon its own personal attributes. It is hoped that this research will help in reducing the cyclist crash and help in the promotion of this mode for the holistic, sustainable, integrated cyclist transportation system. The final output variable, i.e., the trip maker's age group, maybe correlated with many underlying factors. Therefore, future research should aim to create a heterogeneous model, which can uncover the underlying variables.

Acknowledgement: The authors would like to thank the Gateshead City Council for allowing access to the TADU crash database.

Conflict of Interest

We know of no conflicts of interest associated with this publication, and there has been no financial support for this work that could have influenced its outcome.

References

1. Bill E, Rowe D, Ferguson N (2015) Does experience affect perceived risk of cycling hazards?. Scottish Transp. Appl. Res. Conf. 1–19.
2. Wardman M, Hatfield R, Page M (1997) The UK national cycling strategy: Can improved facilities meet the targets?. Transp. Policy 4: 123–133. https://doi.org/10.1016/S0967-070X(97)00011-5.
3. Lawson A, (2015) An Analysis of Cycling Safety and development of a bicycle trip assignment methodology. Dissertation, Trinity College Dublin.
4. Heinen E, Maat K, Van Wee B, (2011) Day-to-Day Choice to Commute or Not by Bicycle. Transp. Res. Rec. J. Transp. Res. Board 2230: 9–18 (2011). https://doi.org/10.3141/2230-02.
5. Sener IN, Eluru N, Bhat CR (2009) An Analysis of Bicycle Route Choice Preferences Using a Web-Based Survey to Examine Bicycle Facilities. Transportation (Amst) 36: 511–539.
6. Guthrie N, Davies DG, Gardner G, (2001) Cyclists ’ assessments of road and traffic conditions : the development of a cyclability index. Transport Research Lab, London, UK.
7. Aldred R, Goodman A (2018) Predictors of the frequency and subjective experience of cycling near misses : Findings from the first two years of the UK Near Miss Project. Accid. Anal. Prev. 110: 161–170. https://doi.org/10.1016/j.aap.2017.09.015.
8. Laureshyn A, Varhelyi A (2018) The Swedish Traffic Conflict Technique: observer’s manual. Lund University.
9. Allen BL, Shin BT (1978) Analysis of Traffic Conflicts and Collisions. Transp. Res. Rec. 667: 67–

74.

10. Schleinitz K, Petzoldt T, Franke-Bartholdt L, Krems J, Gehlert T (2017) The German Naturalistic Cycling Study – Comparing cycling speed of riders of different e-bikes and conventional bicycles. Saf. Sci. 92: 290–297. https://doi.org/10.1016/j.ssci.2015.07.027.
11. Potoglou D, Carlucci F, Cirà A, Restaino M (2018) Factors associated with urban non-fatal road-accident severity. Int. J. Inj. Contr. Saf. Promot. 7300: 1–8. https://doi.org/10.1080/17457300.2018.1431945.
12. Welander G, Ekman R, Svanström L, Schelp L, Karlsson A (1999) Bicycle injuries in Western Sweden: a comparison between counties. Accid. Anal. Prev. 31: 13–19. https://doi.org/10.1016/S0001-4575(98)00040-2.
13. Schepers JP, Heinen E (2013) How does a modal shift from short car trips to cycling affect road safety? Accid. Anal. Prev. 50: 1118–1127 (2013). https://doi.org/10.1016/j.aap.2012.09.004.
14. TRL (2011) Infrastructure and Cyclist Safety: Research Findings TRL Report PPR 580. London, UK.
15. Aldred R, Goodman A, Gulliver J, Woodcock J (2018) Cycling injury risk in London : A case-control study exploring the impact of cycle volumes , motor vehicle volumes , and road characteristics including speed limits. Accid. Anal. Prev. 117: 75–84. https://doi.org/10.1016/j.aap.2018.03.003.
16. Gettman D, Pu L, Sayed T, Shelgy S (2008)Surrogate Safety Assessment Model and Validation : Final Report, Report No. FHWA--HRT-08-051.
17. Kasm OA, Ma Z, Chow JYJ, Diabat A (2019) Quantifying the effect of cyclist behavior on bicycle crashes and fatalities. In: 98th Annual Meeting of the Transportation Research Board. Washington DC, USA.
18. Lovegrove G, Sayed T (2006) Using Macro-level Collision Prediction Models in Road Safety Planning Applications. Transp. Res. Rec. J. Transp. Res. Board 1950: 73–82.
19. Ambros J, Jurewicz C, Turner S, Kiec M (2018) An international review of challenges and opportunity in development and use of crash prediction models. Eur. Transp. Research Rev.
20. Lord D, Mannering F (2010) The statistical analysis of crash-frequency data : A review and assessment of methodological alternatives. Transp. Res. Part A. 44: 291–305. https://doi.org/10.1016/j.tra.2010.02.001.
21. Mannering FL, Bhat CR (2014) Methodological frontier and future directions. Anal. Methods Accid. Res. 1: 1–22. https://doi.org/10.1016/j.amar.2013.09.001.
22. Walker I (2007) Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. Accid. Anal. Prev. 39: 417–425. https://doi.org/10.1016/j.aap.2006.08.010.
23. Department for Transport (2019) Total Road length (miles) by road type and local authority in Great Britain, RDL01012a. London, United Kingdom.
24. Dougherty M (1993)A review of neural networks applied to transport. Transp. Res. Part C. 3: 247–260. https://doi.org/10.1016/0968-090X(95)00009-8.
25. Haykin S (2005) Neural Networks, A Comprehensive Foundation. Pearson Education (Singapore) Pte.Ltd. https://doi.org/10.1142/s0129065794000372.
26. Karlaftis MG, Vlahogianni EI (2011) Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transp. Res. Part C Emerg. Technol. 19: 387–399. https://doi.org/10.1016/j.trc.2010.10.004.
27. Zimmermann HCHJ (1998) Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach. Springer Science plus Business Media, New York, USA. https://doi.org/10.1007/978-94-011-4403-2.
28. Elvik R (2009) The non-linearity of risk and the promotion of environmentally sustainable transport. Accid. Anal. Prev. 41: 849–855. https://doi.org/10.1016/j.aap.2009.04.009.
29. Pallant J (2011) SPSS Survival Manual. Allen & Unwin, Crows Nest, Australia.
30. Environment and Transport Overview and Scrutiny Committee (2015) Road Casualty Reduction Leicestershire 2000 to 2014. London, United Kingdom.
31. Cohen J (1998) Statistical power analaysis for the behavioral science. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
32. Mindell JS, Leslie D, Wardlaw M (2012) Exposure-Based, “Like-for-Like” Assessment of Road Safety by Travel Mode Using Routine Health Data. PLoS One. 7: 1–10. https://doi.org/10.1371/journal.pone.0050606.
33. Gatersleben B, Appleton KM (2007) Contemplating cycling to work; attitudes and perceptions in different stages of change. Transp. Res. Part A. 41: 302–312.

34. AASHTO (2010) Highway Safety Manual. American Association of State Highway and Transportation Officials. Washington, USA.
35. Connors RD, Maher M, Wood A, Mountain L, Ropkins K (2013) Methodology for fitting and updating predictive accident models with trend. Accid. Anal. Prev. 56: 82–94. https://doi.org/10.1016/j.aap.2013.03.009.
36. Greibe P (2003) Accident prediction models for urban roads. Accid. Prev. Prev. 35: 273–285.
37. Jonsson T (2005) Predictive models for accidents on urban links - A focus on vulnerable road users, Lund University.
38. Peltola H, Kulmala R (2010) Accident models. VTT, Technical Research Centre of Finland, Espoo, Finland.
39. Hossain M, Muromachi Y (2009) A Framework for real-time crash prediction : Statistical Approach versus artificial intelligence. Infrastruct. Plan. Rev. Japan Soc. Civ. Eng. 26: 979–988.
40. Lawson AR, Pakrashi V, Ghosh B, Szeto WY (2013) Perception of safety of cyclists in Dublin City. Accid. Anal. Prev. 50: 499–511. https://doi.org/10.1016/j.aap.2012.05.029.
41. Yannis G, Dragomanovits A, Laiou A, Richter T, Ruhl S, Calabretta F (2015) Inventory and critical review of the existing APM's and CMF's and related data sources. Confederation of European Directors of Roads, Brussels, Belgium.
42. Sabey B, Taylor H (1980) The known risk we run: The Highway. Transport Research Lab. Berkshire, UK.
43. Carsten OMJ, Tight MR, Southwell MT, Plows B (1980) Urban accidents: why do they happen. Basingstoke, UK.
44. Calvey JC, Shackleton JP, Taylor MD, Llewellyn R (2015) Engineering condition assessment of cycling infrastructure: Cyclists' perceptions of satisfaction and comfort. Transp. Res. Part A Policy Pract. 78: 134–143. https://doi.org/10.1016/j.tra.2015.04.031.