Research Article

Prediction and Analysis of the Severity and Number of Suburban Accidents Using Logit Model, Factor Analysis and Machine Learning: A case study in a developing country

Meisam Ghasedi[1] · Maryam Sarfjoo[2] · Iraj Bargegol[1]

Received: 5 July 2020 / Accepted: 24 December 2020 / Published online: 7 January 2021

Abstract

The purpose of this study is to investigate and determine the factors affecting vehicle and pedestrian accidents taking place in the busiest suburban highway of Guilan Province located in the north of Iran and provide the most accurate prediction model. Therefore, the effective principal variables and the probability of occurrence of each category of crashes are analyzed and computed utilizing the factor analysis, logit, and Machine Learning approaches simultaneously. This method not only could contribute to achieving the most comprehensive and efficient model to specify the major contributing factor, but also it can provide officials with suggestions to take effective measures with higher precision to lessen accident impacts and improve road safety. Both the factor analysis and logit model show the significant roles of exceeding lawful speed, rainy weather and driver age (30–50) variables in the severity of vehicle accidents. On the other hand, the rainy weather and lighting condition variables as the most contributing factors in pedestrian accidents severity, underline the dominant role of environmental factors in the severity of all vehicle-pedestrian accidents. Moreover, considering both utilized methods, the machine-learning model has higher predictive power in all cases, especially in pedestrian accidents, with 41.6% increase in the predictive power of fatal accidents and 12.4% in whole accidents. Thus, the Artificial Neural Network model is chosen as the superior approach in predicting the number and severity of crashes. Besides, the good performance and validation of the machine learning is proved through performance and sensitivity analysis.

Keywords Predictions · Suburban crashes · Logit model · Factor analysis · Machine learning

1 Introduction

With the growing economy in developing countries, suburban traffic plays a crucial role in the country's comprehensive transportation system. The increase in road transport in comparison to less progress in other types of transportation systems and insufficient infrastructures in Iran, has significantly increased the urban pollution, road users wasted time and above all the damages caused by traffic accidents [1, 2]. The high death rate of traffic accidents in suburban roads is considered as one of the challenging safety issues in developing countries like Iran. According to World Health Organization (WHO), there are more than 20,000 fatalities, and around 300,000 injuries in road traffic accident occurred in Iran each year, which 69% of them belongs to suburban roads' crashes [3, 4]. Therefore, the analysis and investigation of suburban road accidents and providing solutions to reduce them due to local traffic and environmental characteristics are essential to be investigated. It is obvious that such recognition will

✉ Meisam Ghasedi, meisam_ghasedi@msc.guilan.ac.ir; Maryam Sarfjoo, hed.s.kasmaei@gmail.com; Iraj Bargegol, bargegol@guilan.ac.ir | [1]Department of Civil Engineering, FacultyofEngineering, University of Guilan, Rasht, Iran. [2]Department of Environmental Engineering, University of Environment, Karaj, Iran.

SN Applied Sciences (2021) 3:13 | https://doi.org/10.1007/s42452-020-04081-3

lead to the feasibility of developing traffic safety programs of engineers and will enable them to better understand the factors that have a positive or negative impact on the severity of crashes. The ultimate goal of analyzing and studying the data gathered by experts is to reach the most accurate and comprehensive method to forecast type and number of accidents considering the given characteristics such as geographical, physical, and human factors of studied road. Limited available accidents data especially in short-term period or pedestrian crashes is deemed one of the main challenges of engineers. On the other side, dealing with limited number of accidents is the nature of road accidents analysis. Due to numerous preventive and corrective measures utilized by governments, the number of accidents should be reduced as many as possible. Therefore, due to this limitation, it is decided to utilize the statistical approaches (factor analysis and logit model) and machine learning in order to investigate the occurred accidents in one of the busiest suburban highways of Guilan Province located in North of Iran. The final goal is to determine and analyze the most effective parameters on increasing the severity of accidents and present the most accurate prediction model for vehicle and pedestrian accidents separately.

This paper is organized as follows. Section 2 describes the past studies about the application of statistical and artificial neural network approaches in generating and analyzing the prediction model of accidents. Section 3 introduces the study route and the utilized methodology in this paper. Section 4 describes the details of the factor analysis, logit, and Machine Learning approaches and presents the obtained results. Finally, Sect. 5 and 6 demonstrate the differences in results using different modeling methods and present the main conclusions of this study.

2 Previous Studies

A review of past studies in the field of predicting number and severity of crashes indicates that each of them has examined the relationship between effective parameters in accidents with the severity that are classified into different categories [5]. Most of the previous studies in this area have been conducted in two categories of statistical and artificial neural network approaches. Firstly, many researchers had focused on generating models based on a statistical methods to predict the crash numbers. The significant difference in their conducted researches was between the type of model and the number of parameters or independent variables influencing the severity of crashes. Thus, various models of logit or probit had been utilized according to their proportion. In studies that the severity of accidents is divided into two categories, binary logit or probit models have been necessarily utilized, and in studies with more categories of severity, the multiple logit or probit models have been used. Jason and Shanker [6] studied the impact of the fixed roadside objects on the entire urban state route system in Washington State. The utilized models in this research were multivariate nested logit models of injury severity and the severity of collisions were classified into five categories: property damage, minor injury, moderate injury, severe and fatal collisions. The proposed model showed that the utilization of well-designed leading ends of guardrails decreases the number of fatal accidents. The model also indicated the importance of protecting vehicles from collisions with trees stumps and rigid poles that cause severe injury or death. Yan et al. [7] studied the multiple logistic regression model and Quasi-induced exposure concept for rear-end accidents occurring at signalized intersections. In order to study the characteristics of accidents, parameters related to the road environment, striking and struck role were investigated. The most important factors were influencing these types of accidents included number of lanes, divided/undivided highway, accident time, road surface condition, highway character, urban/rural, and speed limit, vehicle type, driver age, alcohol/drug use and driver residence. Deng et al. [8] investigated the severity of head-on collisions in Connecticut State utilizing a sequential probit model. Their studies showed that the wet surface of the pavement and the time of the collision at night are highly correlated with the severity of the collision, while the increase in the width of the lane decreases the severity of the collisions. Kim et al. [9] investigated the severity of bicycle injuries in bicycle–motor vehicle accidents and the factors affecting it. The utilized multinomial logit model could predict the probability of four categories of collisions severity, including fatal, incapacitating, non-incapacitating, and possible or no injury. The results of their modelling showed that a lot of factors such as a truck involving in a collision, high speed, consuming alcohol by driver or cyclist, the age of over 55 years old for the cyclist, inclement weather and head-on collisions lead to an increase in the severity of injuries leading to death. In a study by Peter Savolainen and Fred Mannering in 2007, modeling was utilized once for single-vehicle crashes and once for multi-vehicle crashes, which Nested logit and standard multinomial logit model were used for modeling [10]. The results showed that the parameters such as age, roadway characteristics, alcohol consumption, helmet use, unsafe speed were the most prominent factors which increase the severity of crashes. Pengfei Liu et al. [11] studied the contributing factors that affect the severity of head-on crashes in North Carolina in United States utilizing mixed logit model. Results of their studies maintained that adverse weather condition, two-way divided road, traffic control, young drivers, and pickups would decrease the injury severity of head-on crashes.

The majority of statistical methods have their assumptions and predefined relations between independent and dependent variables, and if these assumptions are violated, the model will provide incorrect prediction of accidents. Machine learning tool seem to be one of the most reliable and efficient approaches dealing with everyday human challenges with the capability of skipping theoretical assumptions. [12] Machine learning approaches using Artificial Neural Network (ANN) could be utilized in various areas such as environment and business sectors to develop prediction models of ambient temperature, energy production and consumption. [13–15] Demirezen et al. proved the competence of artificial neural network (ANN) as a dependable and powerful predicting approach of outdoor temperature with minimum error in two different studies.[16, 17] Banan et al. utilized deep learning neural network as a smart and real-time approach to present an automate identification process of fish species [18]. Fan et al. adopted the multilayer perceptron (MLP) together with spatiotemporal model and the long short-term memory (LSTM) network to make an estimation of temperature distributions during the thermal process. [19] Wu et al. selected ANN to present a rainfall prediction model due to its high efficiency in training large-size samples. [20]

Since crashes are directly related to the human lives, the artificial neural network will have widespread application in making major decisions including prediction of the type and severity of collisions and proposing alternatives in order to reduce it, without the requirement for any predefined assumptions and relations, and with higher accuracy than statistical methods [21, 22]. Nonlinear relationship between variables can be modelled with various types of ANN in order to recognize the effect of influential factors in an event occurred and predict the future events [23–26]. Chang utilized two models of artificial neural network and negative binomial regression for analyzing and modeling road crashes. Comparing these two methods, he concluded that the artificial neural network model is a more accurate and influential method for analyzing freeway accidents [27]. Akgungor and Dogan [28] proposed two models to estimate number of accidents, injuries and fatalities by making us of artificial neural networks and nonlinear regression. Their study showed that the artificial neural network model could present the prediction model with the lowest error. They used the acquired results to evaluate the performance of proposed model for the future of road safety programs in Turkey. In another study in 2009 [29], they presented an artificial neural network and genetic algorithm (GA) model to acquire prediction model for the number of fatal and injury accidents in Ankara, Turkey. The results showed that the artificial neural network model have the least error in training and testing data, resulting in a more reliable and better prediction model for crashes comparing to GA model. Cansız [30] modelled the accidents with the help of Smeed equation and ANN to estimate the number of fatalities in accidents. This study proved their model accuracy and competency of dead prediction’s numbers.The artificial neural network along with log-normal regression models were utilized in a freeway accidents prediction studied by Bagheri et al. [31]. They considered three-year accident data and parameters such as average daily traffic volume, percentage of heavy vehicle, average speed and pavement condition as input variables. At the end of their study, they proved the ANN model efficiency over log-regression model and concluded that the average speed of vehicles and average daily traffic volume are the most influential factors in freeway accidents. Khair et al. [32] predicted crashes that occurred under Jordanian local conditions, utilizing novel artificial neural network model. They asseted that the estimated collisions based on sufficient data were close to the actual number of crashes and thus considered the proposed model reliable for forecasting number of occurred accidents. Afandizadeh et al. [33] started modeling the role of human factors in collisions utilizing the artificial neural network. In this study, they considered accident-prone violations in the suburban highways to select the effective variables in the model designing process. Afterward, they categorized the collision into three levels of severity, property damage, injury, and death, then different structures were built using the artificial neural network, and eventually the model was validated using new data, and the results of the optimal network parameters showed a high accuracy of the neural network in building the model. E. Contreras et al. [34] utilized a model by using ANN to predict traffic accidents in urban zones of Nuevo León city. In this study Scilab development software was used to validate the maximum sensitivity of intended Neural Network. The satisfactory mean square gradient error of the presented model demonstrated the validation of the prediction model.

3 Study route and methodology

In this research, the study route (Chaboksar–Lahijan) is a busy road in the north of Iran, which is known as the most accident-prone suburban highway in the Guilan province. The length of this route is 61 km, and due to lots of accesses, commercial and residential land uses in many parts, especially at the city entrances in which urban texture overcomes suburban texture; therefore, highway traffic performance in this route is challenged. This road is categorized as the most traveled highway in this province, so the accidents' frequency and severity analysis are critical to be investigated. Generally, the whole data consists of 1117 accidents which 56 of them have some deficiencies. Eventually, 1061 accidents (956 vehicle accidents and 105 pedestrian accidents) are obtained for analysis.

In this paper, the dependent variable is the different levels of accident severity, which have been divided into three categories of fatal, injury, and property damage only (PDO) accidents. Since the number of fatal accidents is few compared to total accidents and by considering the three levels of dependent variables, the independent variables significance and goodness-of-fit of a model have not been achieved, therefore, in the case of vehicle accidents, fatal accidents has been merged with accidents leading to injury and the dependent variable is divided into two categories. It should be noted that, in many cases, traffic polices consider the injured persons just in accident scenes; however, the injured may die after being transferred to the hospital or on the way of the hospital; so it leads to an inconsistency in the accidents fatalities statistics. Therefore, merging these two categories is practically sensible, and there is no interference in the study's objective, which is understanding the most effective factors on the severity of accidents. Furthermore, for the analysis of pedestrian accident severity, the dependent variable has been divided into two levels of injury and fatal.

Independent variables affecting the severity of accidents have been categorized for both vehicle and pedestrian accidents according to Table 1. The data should be converted to nominal variables to be used in the modeling process; therefore, all variables have become nominal in a way that number 1 indicates the variable intervention in the accident, and zero indicates the variable non-intervention in the accident. After preparation of data and converting the dependent and independent variables into dummy variables, vehicle and pedestrian accidents will be separately modelled and analyzed using factor analysis, logit and machine learning approaches.

4 Analysis and discussion

4.1 Exploratory factor analysis

In studies with large number of variables, researchers are looking to reduce the number of variables and form a new structure for more practical and accurate data analysis. Therefore, factor analysis is used to identify the principal variables in order to explain the correlation pattern between the observed variables. Factor analysis plays a very important role in identifying hidden variables or factors through observed variables.

The results of Kaiser-Meyer-Olkin (KMO) indexes and the Bartlett tests for vehicle and pedestrian accidents are shown in Table 2. Since the KMO index for vehicle accidents in 2018 and pedestrian accidents are less than 0.5, the factor analysis results would not be reliable for these two mentioned cases. Moreover, the significance value of Bartlett's test for all cases is less than 5%, which rejects the assumption of the known correlation matrix.

The eigenvalues and remaining factors in the analysis should be recognized in order to perform factor analysis. The factors with an eigenvalue of less than one should be excluded from the analysis. Table 3 shows the eigenvalues of sum of three years vehicle accidents occurred between 2017 and 2019.

According to Table 3, factors one to six have an eigenvalue more than one and remain in the analysis. Therefore, Table 4 represent rotated component matrix, which contain estimates of the correlations between each of the variables and the estimated components. The higher coefficients in each row represents the more importance of that variable.

According to factor analysis on the 13 variables affecting the vehicle accidents (2017–2019), six factors are recognized as principal factors. The factor analysis shows that collision with, type of collision and the main cause variables are considered as the first factor affecting the severity of accidents. In addition, the variables of the road surface and weather condition are considered as the second factor. Moreover, accident time and lighting conditions are categorized as the third factor. The at-fault vehicle, age of driver and driver's gender are the fourth factor and road geometric characteristic is regarded as the fifth factor. Finally, the season and day of the accident are considered as the sixth factor. In a nutshell, the importance of "collision with, type of collision and main cause" as the first influential factors on increasing the severity of accidents asserts further attention to details of these sub-variables. The frequency analysis of accidents shows the large share of light vehicle, rear-end and side-impact, lack of attention and driving too close to the car in front in total number of accidents. All of these behaviors are the direct result of careless driving and they are extremely dangerous. They may also result in a serious car crash that has a long-lasting influence on innocent people, drivers, pedestrians, and cyclists alike. Therefore, imposing more penalties such as dramatically increase of insurance rates and driving license suspensions for novice drivers would seem reasonable. It is also suggested to alert inattentive drivers of potential danger by implementing pavement warning methods such as alert strips (sleepy bumps) or installing speed humps, especially at the city entrances along this road in which urban texture overcomes suburban texture.

Table 1 Description of variables used in the study

Variable	Variable levels (vehicle accidents)	Variable levels (pedestrian accidents)
Accident severity	1- PDO	1- Injury
	2- Injury/fatal	2- Fatal
Accident time	1- 00:00 to 06:00	1- 00:00 to 06:00
	2- 06:00 to 12:00	2- 06:00 to 12:00
	3- 12:00 to 18:00	3- 12:00 to 18:00
	4- 18:00 to 24:00	4- 18:00 to 24:00
Accident day	1- Saturday	1- Saturday
	2- Sunday	2- Sunday
	3- Monday	3- Monday
	4- Tuesday	4- Tuesday
	5- Wednesday	5- Wednesday
	6- Thursday	6- Thursday
	7- Friday	7- Friday
Season	1- Spring	1- Spring
	2- Summer	2- Summer
	3- Autumn	3- Autumn
	4- Winter	4- Winter
Weather	1- Sunny	1- Sunny
	2- Cloudy	2- Cloudy
	3- Rainy	3- Rainy
	4- Snowy	4- Snowy
	5- Foggy	5- Foggy
	6- Stormy	6- Stormy
Road surface condition	1- Dry	1- Dry
	2- Humid	2- Humid
Geometry of accident location	1- Straight Section	–
	2- Horizontal Curve	–
Lighting condition	1- Day	1- Day
	2- Night (Adequate lighting)	2- Night (Adequate lighting)
	3- Night (Inadequate lighting)	3- Night (Inadequate lighting)
Collision with	1- Light vehicle	–
	2- Heavy vehicle	–
	3- Motorcycle	–
	4- Curb and fixed object	–
	5- Animal	–
	6- Unknown object	–
At-fault Vehicle	1- Light vehicle	1- Light vehicle
	2- Pickup truck	2- Pickup truck
	3- Heavy vehicle	3- Heavy vehicle
	4- Motorcycle	4- Motorcycle
	5- Fleeing vehicle and miscellaneous	5- Fleeing vehicle and miscellaneous
	6- Unknown object	–
Type of collision	1- Head-on collision	–
	2- Rear-end collision	–
	3- Side-impact collision	–
	4- Sideswipe collision	–
	5- Rollover	–
Driver gender	1- Male	1- Male
	2- Female	2- Female

Table 1 (continued)

Variable	Variable levels (vehicle accidents)	Variable levels (pedestrian accidents)
Driver age	1- Less than 18 yrs	1- Less than 18 yrs
	2- 18–30 yrs	2- 18–30 yrs
	3- 30–40 yrs	3- 30–40 yrs
	4- 40–50 yrs	4- 40–50 yrs
	5- 50–60 yrs	5- 50–60 yrs
	6- 60–70 yrs	6- 60–70 yrs
	7- 70–86 yrs	7- 70–86 yrs
Reason of accident	1- Failure to yield the right-of-way	1- Lack of attention
	2- Lack of attention	2- Backover movement
	3- Yaw motion of the vehicle to the left	3- Unsafe Lane Changes
	4- Exceeding lawful speed	4- Wrong way movements
	5 Failure to observe longitudinal spacing	5- Inability to control the vehicle
	6- Backover movement	–
	7– Unsafe Lane Changes	–
	8– Wrong way movements	–
	9– Inability to control the vehicle	–
	10– Improper Turns	–
	11– Sudden door opening of vehicle	–
	12– Technical defect in the vehicle	–

Table 2 KMO and Bartlett tests

Year		2016 (Vehicle Accidents)	2017 (Vehicle Accidents)	2018 (Vehicle Accidents)	Sum of 3 years (Vehicle Accidents)	Sum of 3 years (Pedestrian Accidents)
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		0.554	0.5	0.493	0.525	0.487
Bartlett's Test of Sphericity	Approx Chi–Square	731.236	225.534	310.278	957.416	78.042
	DF	66	66	66	66	46
	Significance	0.000	0.000	0.000	0.000	0.002

Due to the large number of rainy days in this highway, and considering the surface, weather, time and lighting condition as second and third most influential factors, both these mentioned issues indicate the importance of implementing corrective actions such as increasing highway lighting condition and improving pavement surface quality in addition to preventive measures such as installing more VMS (Variable Message Signs) and traffic speed cameras especially in bad weather conditions.

4.2 Modeling using logit model

In order to analyze data and obtain a prediction model using logit model, there are three ways to enter variables, including Entering, Backward and Forward approaches. Since all variables are entered simultaneously into the equation in the first method (Enter), this model does not have the opportunity to process data appropriately and extract the most significant variables, so it cannot be a suitable method. Therefore, the Backward and Forward methods are used to enter data into the logit equation. The one with higher accuracy in predicting the number of accidents will be recognized as the superior method. Table 5 summarizes logit models in forward and backward methods. As it is mentioned earlier, prediction accuracy determines the superior model, so the Backward method with a higher percentage of accurate predictions is chosen as the best method for making models in all cases. Besides, Tables 6 indicates the chi-square, degree of freedom (df), and significance (sig) of the Backward method in the modeling process. Since the significance of two backward models used to predict vehicles and pedestrians' accidents are zero, the capability of the model to predict accidents is confirmed.

Table 3 Eigenvalues of vehicle accidents in (2017–2019)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.805	13.885	13.885	1.805	13.885	13.885	1.748	13.444	13.444
2	1.533	11.794	25.679	1.533	11.794	25.679	1.508	11.600	25.044
3	1.351	10.396	36.075	1.351	10.396	36.075	1.363	10.487	35.530
4	1.204	9.258	45.332	1.204	9.258	45.332	1.228	9.444	44.974
5	1.024	7.877	53.210	1.024	7.877	53.210	1.045	8.039	53.013
6	1.006	7.740	60.950	1.006	7.740	60.950	1.032	7.937	60.950
7	0.944	7.263	68.213	–	–	–	–	–	–
8	0.880	6.768	74.981	–	–	–	–	–	–
9	0.872	6.708	81.690	–	–	–	–	–	–
10	0.828	6.372	88.061	–	–	–	–	–	–
11	0.632	4.858	92.920	–	–	–	–	–	–
12	0.550	4.233	97.153	–	–	–	–	–	–
13	0.370	2.847	100.000	–	–	–	–	–	–

Extraction Method: Principal Component Analysis

Table 4 Rotated component matrix for vehicle accidents (2017–2019)

	Component					
	1	2	3	4	5	6
Time	– 0.133	– 0.194	0.796	0.005	– 0.065	– 0.041
Season	0.088	0.331	– 0.055	0.230	0.311	0.443
Day	– 0.039	– 0.110	0.065	– 0.081	– 0.106	0.895
Surface	– 0.033	0.831	0.030	– 0.020	– 0.087	0.004
Geometry	0.101	0.010	0.017	– 0.047	0.882	– 0.039
Lighting Condition	0.073	0.166	0.823	0.012	0.067	0.085
Collision With	0.865	– 0.022	0.015	0.040	0.005	0.029
At–fault Vehicle	0.174	– 0.028	– 0.107	0.656	– 0.205	0.049
Age	– 0.202	– 0.094	– 0.064	0.566	0.256	– 0.023
Sex	0.063	– 0.111	– 0.156	– 0.636	0.023	0.028
Type of collision	0.847	0.058	– 0.030	0.015	– 0.075	– 0.119
Weather	0.010	0.777	– 0.037	0.019	0.089	– 0.039
Main Cause	0.405	– 0.027	– 0.036	– 0.086	0.145	0.061

Extraction Method: Principal Component Analysis

Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 5 iterations

Table 5 Prediction accuracy of regression models

Regression type	Vehicle accidents (2017)	Vehicle accidents (2018)	Vehicle accidents (2019)	Vehicle accidents (sum of 3 years)	Pedestrian accidents (sum of 3 years)
Forward	78.2%	74.3%	82.7%	75.8%	77.1%
Backward	79.9%	79.6%	84%	78.2%	80%

Tables 7 and 8 indicate the effective variables on making a prediction model for both vehicles and pedestrian accidents. Since the factor analysis specified collision with, type of collision and the main cause variables as the first factor affecting the severity of accidents, the logit model proved this result. According to Table 7, after "collision with" variables, the most effective variables increasing the severity of vehicle crashes are respectively exceeding

Table 6 Backward model coefficients (vehicle and pedestrian accidents)

	Final Step	Chi–square	df	Sig
Accidents of Vehicles (Sum of 3 years)				
Step 33	Step	−2.254	1	0.133
	Block	446.006	29	0.000
	Model	446.006	29	0.000
Accidents of Pedestrians (Sum of 3 years)				
Step 23	Step	−0.533	1	0.465
	Block	137.32	9	0.000
	Model	137.32	9	0.000

lawful speed, rainy weather, driver age (30–40), driver age (40–50). In addition to implementing corrective and preventive actions mentioned in factor analysis sector about rainy weather condition and exceeding lawful speed due to poor visibility and violation of speed limit, this result asserts the role of drivers at the age between 30 and 50 years on the rise of the severity of accidents. It could be related to the tendency of these drivers to higher speeds considering their more skills at this age range. It clearly shows that the new drivers and also the more experienced drivers with age of older than 50 are more cautious in driving. Therefore, government should provide more applicable education by focusing on this age group to warn them about careless driving behaviors.

According to the variables' coefficients for pedestrian accidents (Table 8) through the logit model, the three most effective variables influencing the severity of pedestrian crashes are respectively rainy weather, heavy vehicle and lighting condition. The repetition of rainy weather and lighting condition as effective variables on the severity of pedestrians accident, maintains the significant role of these factors.

4.3 Modeling using artificial neural network

Several types of neural networks can be used to make an artificial neural network prediction model. Considering that the qualitative data used in this study, a neural network with pattern recognition capability is used to make the prediction model. Pattern recognition is an

Table 7 Variables of the severity of vehicle accident logit model in 2017, 2018, and 2019

Predictive variables	(β)	Standard deviation	Wald statistic	Significance	Odds ratio
Season (Spring)	0.531	0.194	7.482	0.006	1.7
Season (Fall)	0.422	0.232	3.305	0.069	1.525
Day of the accident: Thursday	−0.4	0.214	3.513	0.061	0.67
Collision with light vehicle	4.417	0.55	64.451	0.000	82.828
Collision with heavy vehicle	4.264	0.668	40.703	0.000	71.059
Collision with curb and fixed object	3.312	0.611	29.411	0.000	27.448
Collision with unknown object	2.567	0.678	14.343	0.000	13.029
At–fault vehicle (pickup truck)	0.501	0.294	2.895	0.089	1.65
At–fault vehicle (motorcycle)	−3.626	0.540	45.104	0.000	0.027
Age (18–30 years)	0.928	0.463	4.018	0.045	2.528
Age (30–40 years)	1.376	0.470	8.57	0.003	3.959
Age (40–50 years)	1.445	0.479	9.093	0.003	4.240
Age (50–60 years)	0.859	0.526	2.671	0.102	2.362
Head–on collision	−1.062	0.343	9.577	0.002	0.346
Rear–end collision	−0.587	0.242	5.884	0.015	0.556
Side–impact collision	−1.109	0.263	17.815	0.000	0.330
Weather (Rainy)	1.52	0.846	3.233	0.072	4.573
Weather (Snowy)	1.078	0.372	8.388	0.004	2.934
The main cause (failure to yield the right–of–way)	−1.249	0.367	11.559	0.001	0.287
The main cause (lack of attention)	−0.781	0.335	5.42	0.020	0.458
Yaw motion of the vehicle to the left	−1.068	0.621	2.954	0.086	0.344
Exceeding lawful speed	1.59	0.859	3.428	0.064	0.204
Unsafe Lane Changes	−0.952	0.382	6.223	0.013	0.336
Inability to control the vehicle	−1.314	0.413	10.121	0.001	0.269
Constant	12.938	11,395.1	000/0	0.999	415,541.2

Table 8 Variables of the severity of pedestrian accidents logit model in 2017, 2018, and 2019

Predictive variables	(β)	SD	Wald statistic	Significance	OR
Season (summer)	−1.038	0.596	3.032	0.082	0.354
Day of the accident (wednesday)	1.743	1.055	2.73	0.098	5.713
Day of the accident (thursday)	−1.296	0.668	3.763	0.052	0.274
Lighting condition (adequate light)	−1.917	0.773	6.154	0.013	0.147
At–fault vehicle (heavy vehicle)	−2.045	1.165	3.08	0.079	0.129
Weather (rainy)	2.43	0.645	13.66	0.000	12.39
Constant	−63.374	80,974.516	0.000	0.999	0.000

important component of neural network applications in computer vision, radar processing, speech recognition, and text classification. It works by classifying input data into objects or classes based on key features, using either supervised or unsupervised classification.

The input attributes and output labels used in the machine learning approach are the same as the mentioned variables in Table 1. It is worth mentioning that, as it is clarified in "study route and methodology" section, the dependent variable (output class) is the different levels of accident severity. It has been divided into two categories of fatal/injury, and property damage only (PDO) for vehicle accidents, and two levels of injury and fatal for pedestrian accidents. Then it is time to build a neural network by software. In this study, the used ANN is an application of an existing algorithm. The neural network's input data is divided into three categories:

- Training: These are presented to the network during training for learning process, and the network is adjusted according to its error.
- Validation: These are used to measure network generalization, and to halt training when generalization stops improving.
- Testing: These have no effect on training and so provide an independent measure of network performance during and after training. In other words, it is the main criterion to realize how much the neural network's findings are similar to the actual result.

The details of the accident data entry to the software and its Mean Squared Error and Percent Error are shown in Table 9. Since the number of occurred accidents separately between 2017 and 2019, as well as the sum of three years, is sufficient for the network training process, 70% of the data is used for network training and 15% of the data is used for validation process and the remaining 15% is considered as a test of the built network. Furthermore, due to

Table 9 Details of data entry

Year	Number of Dummy Inputs	Number of Hidden Layers	Number of Outputs	Number of Training–Validation–Testing Samples			
					Samples	MSE	%E
2017	66	10	2	Training	338	$1.02893e^{-1}$	$13.01775e^{-0}$
				Validation	72	$1.49935e^{-1}$	$20.83333e^{-0}$
				Testing	72	$1.73922e^{-1}$	$25.00000e^{-0}$
2018	66	11	2	Training	133	$1.000096e^{-1}$	$10.52631e^{-0}$
				Validation	17	$1.79309e^{-1}$	$35.29411e^{-0}$
				Testing	17	$2.008183e^{-1}$	$29.41176e^{-0}$
2019	66	13	2	Training	215	$5.10555e^{-2}$	$5.11627e^{-0}$
				Validation	46	$1.77010e^{-1}$	$23.91304e^{-0}$
				Testing	46	$1.57805e^{-1}$	$21.73913e^{-0}$
Sum of 3 Years for Vehicle Accidents	66	11	2	Training	670	$1.08855e^{-1}$	$13.58208e^{-0}$
				Validation	143	$1.54477e^{-1}$	$24.47552e^{-0}$
				Testing	143	$1.83499e^{-1}$	$25.87412e^{-0}$
Sum of 3 Years for Pedestrian Accidents	46	10	2	Training	83	$1.58403e^{-1}$	$24.09638e^{-0}$
				Validation	11	$1.95070e^{-1}$	$27.27272e^{-0}$
				Testing	11	$1.20227e^{-1}$	$9.09090e^{-0}$

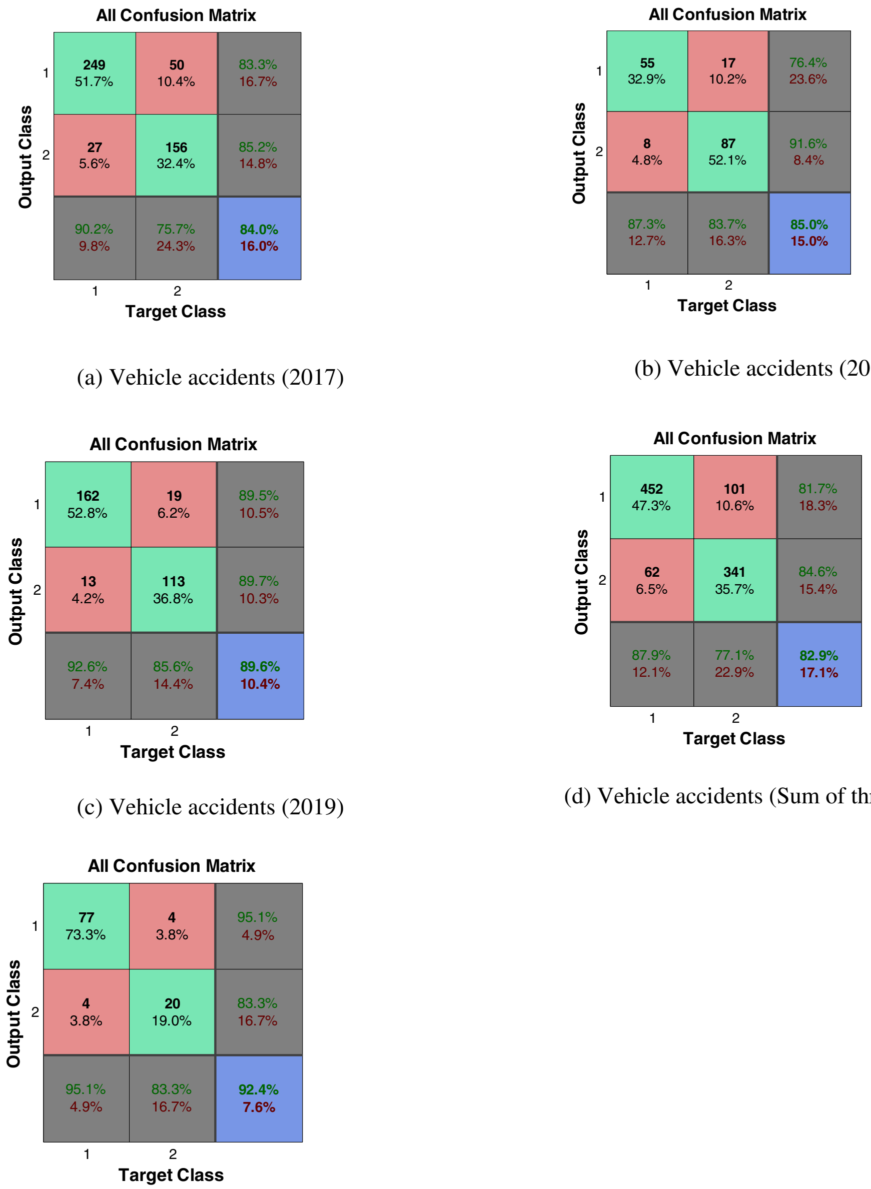
fewer pedestrian occurred accidents during three years, 80% of data is used for network training, 10% as validation, and the remaining 10% as testing.

4.3.1 Results of confusion matrix

Figure 1 indicates the confusion matrix of three modes of training, testing and validation of the created neural network of vehicle-pedestrian accidents. This matrix helps to show the accuracy of the network in the prediction of accidents (PDO, injury and fatal). The squares (1.1) and (2.2) indicated in green squares are the cases which correctly predicted by the network and the squares (1, 2) and (2, 1) indicated in red squares are the cases which present an false prediction of the network. Finally, the blue square shows the total predictive power of the network.

As an illustration, Fig. 1d demonstrates the confusion matrix of vehicle accidents for sum of three years. According to the this matrix, which represents the result of the three processes of training, validation and testing of the network, out of 514 property-damage accidents, 452 cases, and out of 442 injury/fatal accidents, 341 cases are predicted correctly by the model. The prediction accuracy of property-damage accidents in the model is 87.9% and the prediction accuracy of injury/fatal accidents is 77.1%. For a more accurate explanation of the squares of the matrix, the square (1.1) indicates that 452 accidents are correctly predicted as PDO and square (1.2) denotes

Fig. 1 Confusion matrix of accidents



(a) Vehicle accidents (2017)

(b) Vehicle accidents (2018)

(c) Vehicle accidents (2019)

(d) Vehicle accidents (Sum of three years)

(e) Pedestrian accidents (Sum of three years) that 101 accidents leading to fatality or injury are wrongly predicted as PDO. In addition, square (2.1) indicates that 62 fatal or injury accidents are also mistakenly predicted as PDO and square (2.2) suggests that 341 accidents are correctly predicted as fatal or injury accidents. Finally, the blue square represents the overall vehicle accidents predictive power of the network is 82.9%.

4.3.2 The results of the performance of neural network

Figure 2 indicates the performance of neural network training process of vehicle and pedestrian accidents. The indicated circle on these figures shows that since that point on, the answers do not improve and after repeating the process to a given value, which has specified in the horizontal axis, the training process has stopped. The mentioned point is the compromise point with specific mean squared error indicates the best point for the completion of the calculation and the creation of an artificial neural network for the given data.

4.3.3 Sensitivity and specificity analysis of the neural network for the given accident data

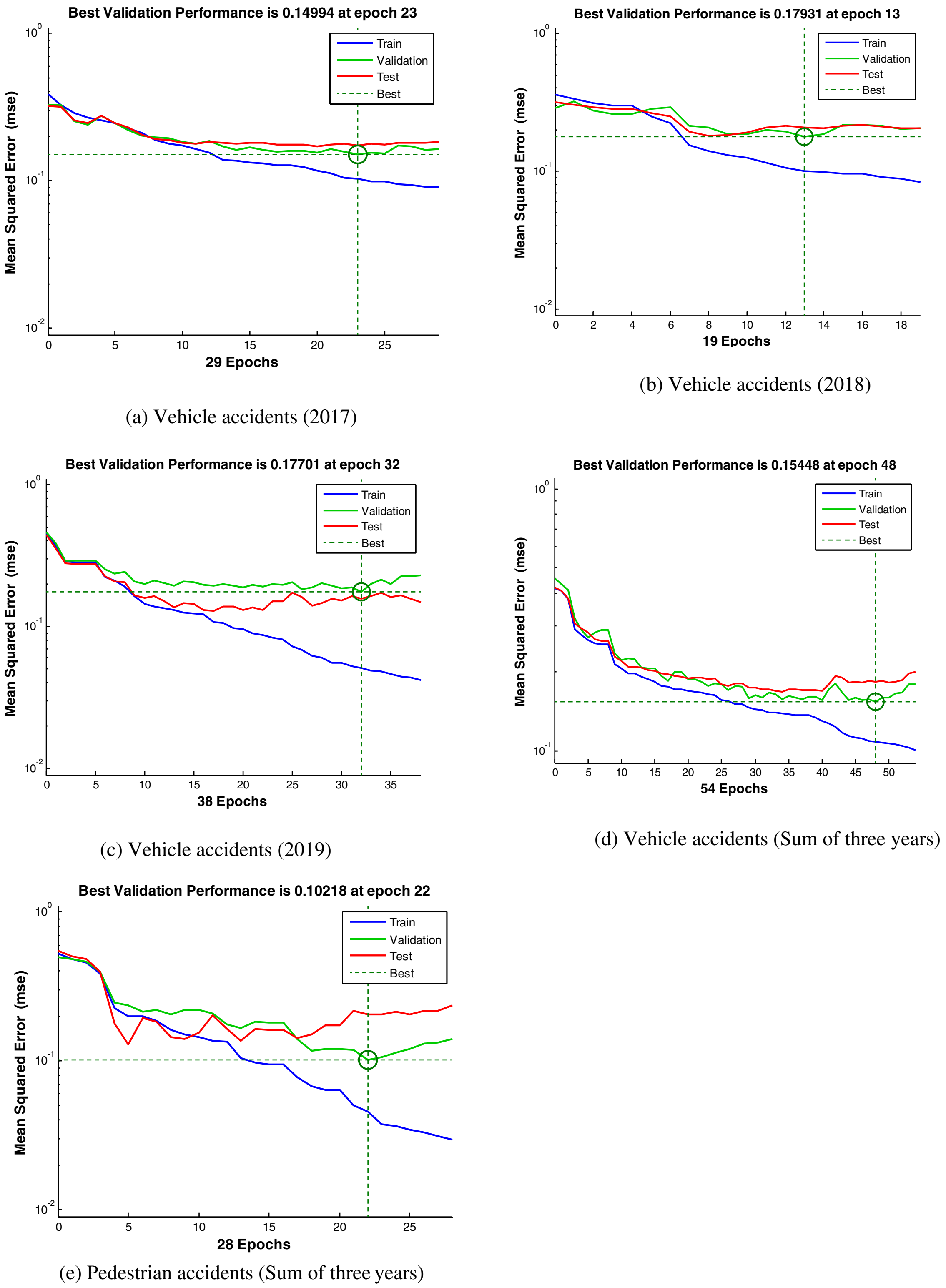
The Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate (sensitivity) versus the false positive rate as the threshold is varied. A perfect test would show points in the upper-left corner, with 100% sensitivity and 100% specificity. Figure 3 analyzes the sensitivity of the network to the correct prediction for vehicle-pedestrian accidents. Class 1 on the diagram indicates network accuracy for existing accidents and class 2 indicates the accuracy of network prediction for future accidents. As the curve goes more upper-left corner, the network is more powerful to predict and estimate correctly the answers which perform very well in this network for both passenger and vehicle accidents.

5 Discussion on the differences in results using different modeling methods

Using factor analysis and logit model simultaneously could contribute to achieving the most comprehensive and efficient model to specify the major contributing factors and their effects on accidents. This being the case, using these results together with the ANN approach as a strong predictive solution provide officials with suggestions to take effective measures to lessen accident impacts and improve road safety. According to Fig. 4, considering two types of accidents leading to damage and the ones leading to death or injury in the vehicle accidents and the accidents leading to death or injury in pedestrian accidents, the artificial neural network modelling has higher accuracy than the statistical methods such as Logit. Therefore, the machine learning model's competence in accident modelling and prediction has been proved as one of the metaheuristic methods compared to statistical methods.

6 Conclusions

In this study, data of accidents in one of the suburban and most accident-prone highways of Guilan province in the north of Iran were collected from 2017 to 2019, including both vehicle and pedestrian accidents. Firstly, the factor analysis was utilized to obtain the classification of the most significant factors affecting the severity of accidents. The combined result of factor analysis and logit model proved the substantial roles of exceeding lawful speed, rainy weather, driver age (30–50) variables in the severity of vehicle accidents. The repetition of rainy weather and lighting condition as influential variables on the severity of pedestrians accident asserted the significant roles of these factors in the whole accident numbers. Thus, the officials should pay more attention to corrective measures such as increasing highway lighting condition and improving pavement surface quality in addition to preventive measures such as installing more VMS (Variable Message Signs) and traffic speed cameras, especially in bad weather conditions. Finally, machine learning was used to build a prediction model of accidents. Comparing the accuracy of logit and the utilized ANN model in this study, the results showed that the machine learning as a metaheuristic approach could lead to better prediction power, particularly in pedestrians' accident in all cases. By using this approach, the effect of corrective measures in reducing the number and severity of accidents would be predictable with high precision. For future studies, it is suggested that other significant data relevant to accidents, including pedestrian clothing color, traffic characteristics of the road, and geographic coordinates of the accident location, would be considered for analyzing the accidents. Implementing these valuable data by Geographical Information System alongside statistical analysis and machine-learning approaches could provide valuable information for a more precise and comprehensive analysis of accident severity.



(a) Vehicle accidents (2017)

(b) Vehicle accidents (2018)

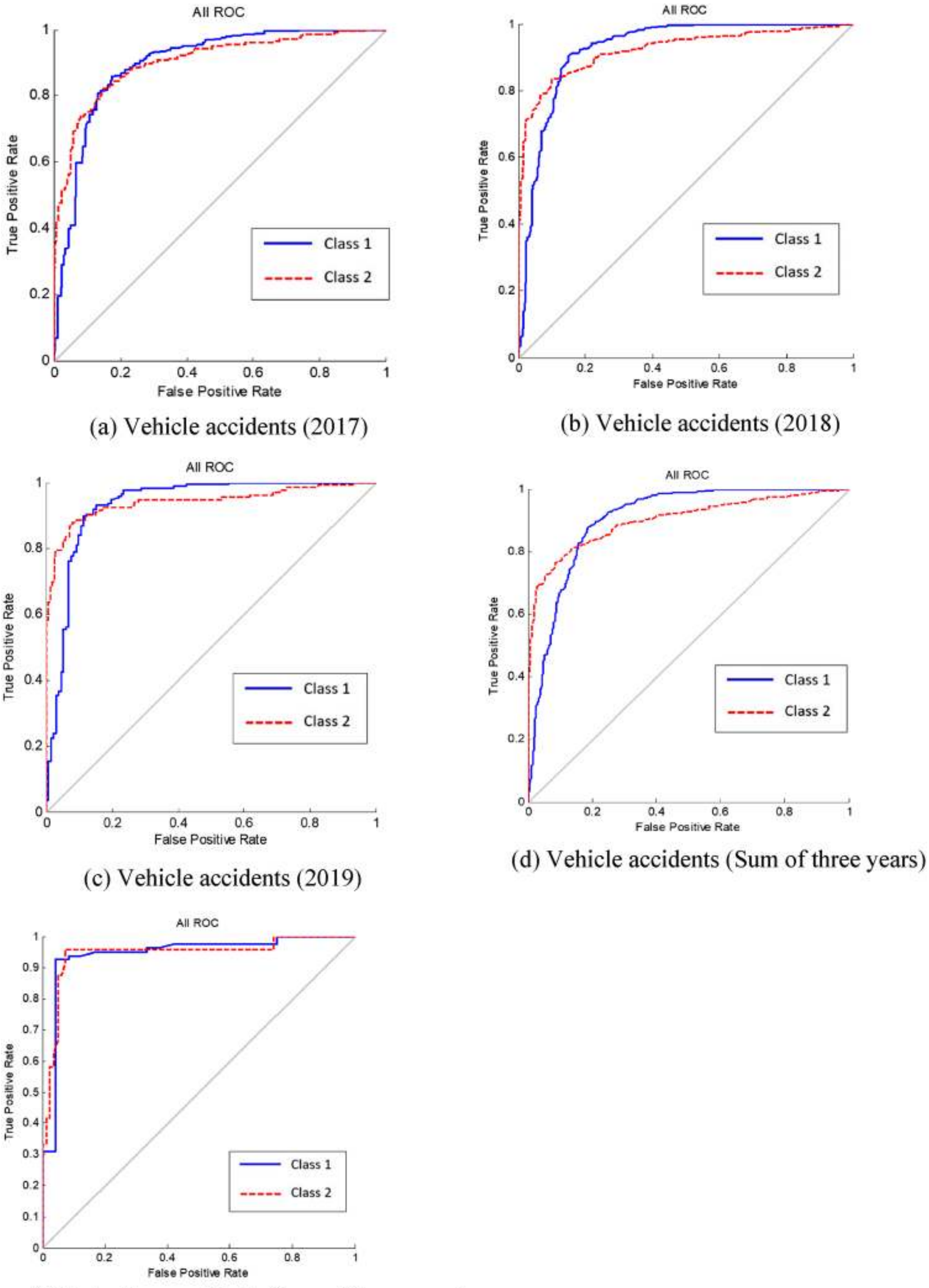
(c) Vehicle accidents (2019)

(d) Vehicle accidents (Sum of three years)

(e) Pedestrian accidents (Sum of three years)

Fig. 2 Performance of neural network training process

Fig. 3 Sensitivity and specificity



(a) Vehicle accidents (2017)

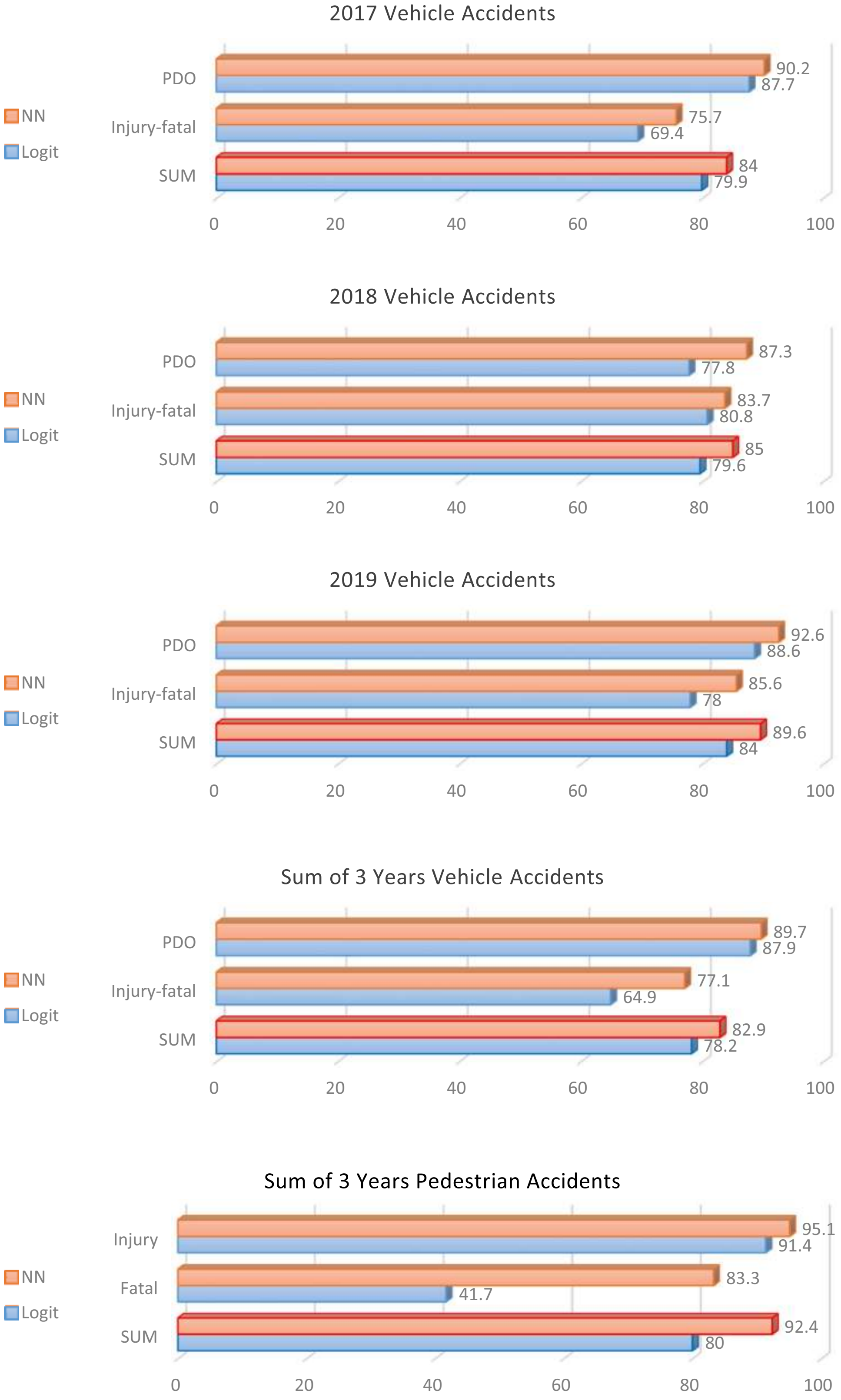
(b) Vehicle accidents (2018)

(c) Vehicle accidents (2019)

(d) Vehicle accidents (Sum of three years)

(e) Pedestrian accidents (Sum of three years)

Fig. 4 Comparison of modeling accuracy by logit method and artificial neural network



Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Consent for publication The authors declare that the contents of this article have not been published previously. All the authors have contributed to the work described, read and approved the contents for publication in this journal.

Human or animal rights All the authors have been certified by their respective organizations for human subject research.

References

1. Zimmerman K, Jinadasa D, Maegga B, Guerrero A (2015) Road traffic injury on rural roads in Tanzania: measuring the effectiveness of a road safety program. Traffic Inj Prev 16:456–460
2. Bargegol I, Gilani VNM, Ghasedi M, Ghorbanzadeh M (2016) Delay modeling of un-signalized roundabouts using neural network and regression. Comput Res Prog Appl Sci Eng (CRPASE) 2:28–34
3. World Health Organization. Global Plan for the Decade of Action for Road Safety 2011–2020. http://www.who.int/roadsafety/decade_of_action/plan/plan_english.pdf
4. Ziari H, Amini A, Saadatjoo A, Hosseini SM, Gilani VNM (2017) A prioritization model for the immunization of accident prone using multi-criteria decision methods and fuzzy hierarchy algorithm. Comput Res Prog Appl Sci Eng (CRPASE) 3:123–131
5. Lee JY, Chung JH, Son B (2008) Analysis of traffic accident size for Korean highway using structural equation models. Accid Anal Prev 40:1955–1963. https://doi.org/10.1016/j.aap.2008.08.006
6. Jason MH, Shankar V, Ulfarsson GF (2005) The crash severity impacts of fixed road side objects. J Saf Res 36:139–147. https://doi.org/10.1016/j.jsr.2004.12.005
7. Yan X, Radwan E, Abdel Aty M (2005) Characteristics of rear-end accident at signalized intersection using multiple logestic regression model. Accid Anal Prev 37:983–995. https://doi.org/10.1016/j.aap.2005.05.001
8. Deng Z, Ivan JN, Garder P (2006) Analysis of factors affecting the severity of head-on crashes: two-lane rural highways in connecticut. Transp Res Record J Transp Res Board 1953:137–146. https://doi.org/10.1177/0361198106195300116
9. Ki Kim J, Kim S, Ilfarsson GF, Porrello LA (2007) Bicycle injury severities in bicycle—motor vehicle accidents. Accid Anal Predict 39:238–251. https://doi.org/10.1016/j.aap.2006.07.002
10. Savolainen P, Manneting F (2007) Probabilistic models of motorcyclists. Injury severities in single—an multi—vehicle crashes. Accid Anal Prev 39:995–965. https://doi.org/10.1016/j.aap.2006.12.016
11. Liu P, Fan WD (2019) Modeling head-on crash severity on NCDOT freeways: a mixed logit model approach. Can J Civil Eng 46:322–328. https://doi.org/10.1139/cjce-2018-0262
12. Giannini F, Laveglia V, Rossi A, Zanca D, Zugarini A (2017) Neural networks for beginners. A fast implementation in matlab, torch, tensorflow. arXiv preprint
13. Shamshirband S, Rabczuk T, Chau K (2019) A survey of deep learning techniques: application in wind and solar energy resources. IEEE Access 7(1):164650–164666. https://doi.org/10.1109/ACCESS.2019.2951750
14. Ardabili S, Najafi B, Shamshirband S, Minaei B, Chand Deo R, Chau K (2018) Computational intelligence approach for modeling hydrogen production: a review. Eng Appl Comput Fluid Mech 12(1):438–458. https://doi.org/10.1080/19942060.2018.1452296
15. Ye KK, Demirezen G, Fung AS, Janssen E (2020) The use of artificial neural networks (ANN) in the prediction of energy consumption of air-source heat pump in retrofit residential housing. E&ES 463(1):012165. https://doi.org/10.1088/1755-1315/463/1/012165/meta
16. Demirezen G, Fung AS (2019) Application of artificial neural network in the prediction of ambient temperature for a cloud-based smart dual fuel switching system. Energy Procedia 158:3070–3075. https://doi.org/10.1016/j.egypro.2019.01.992
17. Demirezen G, Fung AS, Deprez M (2020) Development and optimization of artificial neural network algorithms for the prediction of building specific local temperature for HVAC control. Int J Energy Res. https://doi.org/10.1002/er.5537
18. Banan A, Nasiri A, Taheri-Garavand A (2020) Deep learning-based appearance features extraction for automated carp species identification. Aquacult Eng 89:102053. https://doi.org/10.1016/j.aquaeng.2020.102053
19. Fan Y, Xu K, Wu H, Zheng Y, Tao B (2020) Spatiotemporal modeling for nonlinear distributed thermal processes based on KL decomposition, MLP and LSTM network. IEEE Access 8:25111–25121. https://doi.org/10.1109/ACCESS.2020.2970836
20. Wu CL, Chau KW (2013) Prediction of rainfall time series using modular soft computing methods. Eng Appl Artif Intell 26(3):997–1007. https://doi.org/10.1016/j.engappai.2012.05.023
21. Kamboozia N, Ziari H, Behbahani H (2018) Artificial neural networks approach to predicting rut depth of asphalt concrete by using of visco-elastic parameters. Constr Build Mater 158:873–882. https://doi.org/10.1016/j.conbuildmat.2017.10.088
22. Behbahani H, Ziari H, Amini A, Gilani VNM, Salehfard R (2016) Investigation of un-signalized roundabouts delay with adaptive-network-based fuzzy inference system and fuzzy logic. Comput Res Prog Appl Sci Eng (CRPASE) 2:140–149
23. de Oña J, Mujalli RO, Calvo FJ (2011) Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid Anal Prev 43:402–411. https://doi.org/10.1016/j.aap.2010.09.010
24. Pashaei A, Ghatee M, Sajedi H (2019) Convolution neural network joint with mixture of extreme learning machines for feature extraction and classification of accident images. J Real-Time Image Proc. https://doi.org/10.1007/s11554-019-00852-3
25. Zhu W, Miao J, Hu J, Qing L (2014) Vehicle detection in driving simulation using extreme learning machine. Neurocomputing 128:160–165. https://doi.org/10.1016/j.neucom.2013.05.052
26. Yu B, Chen Y, Bao S, Xu D (2018) Quantifying drivers' visual perception to analyze accident-prone locations on two-lane mountain highways. Accid Anal Prev 119:122–130. https://doi.org/10.1016/j.aap.2018.07.014

27. Chang L (2005) Analysis of highway accident frequencies: Negative, binomial regression versus artificial neural network. Safety Sci 43:541–557. https://doi.org/10.1016/j.ssci.2005.04.004
28. Akgungor A, Dogan E (2008) Estimating road accidents of turkey based on regression analysis and artificial neural network approach, data and information technology; highways; planning and forecasting; safety and human factors. https://trid.trb.org/view/873702
29. Akgungor A, Dogan E (2009) An artificial intelligent approach to traffic accident estimation: model development and application. Transport 24:135–142. https://doi.org/10.3846/1648-4142.2009.24.135-142
30. Cansız OF (2011) Improvements in estimating a fatal accidents model formed by an artificial neural network. SIMULATION 87:512–522. https://doi.org/10.1177/0037549710370842
31. Fatemeh BK, Abdolreza S, Abbas M (2012) Variable efficiency appraisal in freeway accidents using artificial neural networks—case study, cictp, multimodal transportation systems-convenient, safe, cost-effective, efficient, pp. 2657–2664. https://doi.org/10.1061/9780784412442.269
32. Khair SJ, Muaath AF, Hala FG (2014) Prediction of road traffic accidents in jordan using artificial neural network (ANN). J Traf Logist Eng 2:92–94
33. Afandi Z, Nabizadeh MS, Javadi N (2014) Modeling the role of human factor in accidents using artificial neural network. In: 14th International Conference on Transport and Traffic
34. Contreras IE, Torres-Treviño L, Torres F (2018) Prediction of car accidents using a maximum sensitivity neural network. In: Torres GF, Lozoya-Santos J, Gonzalez Mendivil E, Neira-Tovar L, Ramírez Flores P, Martin-Gutierrez J (eds) Smart technology lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 213. Springer, Cham

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.