



OPEN

# Application of Extremely Randomised Trees for exploring influential factors on variant crash severity data

Farshid Afshar<sup>1</sup>, Seyedehsan Seyedabrishami<sup>1✉</sup> & Sara Moridpour<sup>2</sup>

Crash severity models play a crucial role in evaluating the influencing factors in the severity of traffic crashes. In this study, Extremely Randomised Tree (ERT) is used as a machine learning technique to analyse the severity of crashes. The crash data in the province of Khorasan Razavi, Iran, for a period of 5 years from 2013 to 2017, is used for crash severity model development. The dataset includes traffic-related variables, vehicle specifications, vehicle movement, land use characteristics, temporal characteristics, and environmental variables. In this paper, Feature Importance Analysis (FIA), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) plots are utilised to analyse and interpret the results. According to the results, the involvement of vulnerable road users such as motorcyclists and pedestrians alongside traffic-related variables are among the most significant variables in crash severity. Results show that the presence of motorcycles can increase the probability of injury crashes by around 30% and almost double the probability of fatal crashes. Analysing the interaction of PDPs shows that driving speeds above 60 km/h in residential areas raises the probability of injury crashes by about 10%. In addition, at speeds higher than 70 km/h, the presence of pedestrians approximately increases the probability of fatal crashes by 6%.

In recent years, injuries and casualties resulting from road crashes have been one of the leading death causes in Iran and stood as the third major death cause between 2005 and 2017. During an 8-year period from 2013 to 2020, more than 2 million people were seriously injured, and nearly 100,000 people died in traffic crashes in Iran<sup>1</sup>. The large number of crash fatalities and severe injuries also impose a significant economic burden. According to the<sup>2</sup>, in 2016, the cost of fatalities and serious injuries in Iran was around \$28.5 billion accounts for 6.8% of the country's Gross Domestic Product (GDP).

Khorasan-Razavi, in the northeast of the country, is the second most populated province in Iran and is chosen by many people as a leisure trip destination every year. Khorasan-Razavi has one of the highest crash casualties on rural roads across provinces. Between 2013 and 2020, 7040 fatalities and 165,339 severe injuries were observed on rural roads in this province, which both were the second-highest in the country. These statistics suggest a need for an in-depth understanding of the key factors contributing to crash severity in rural areas. However, there are few studies in which real-time traffic data is used to analyse the severity of crashes in rural areas in developing countries.

According to the literature, two approaches are available to identify the influencing factors on crash severity, including statistical models and data-driven methods. Statistical models are the primary method in crash frequency and severity analysis. These models have specific assumptions about model configuration and the used datasets. For instance, there is a linear additive assumption regarding the relationship between independent and dependent variables, which may not be completely accurate in some cases. This functional form can have a negative impact on results due to outliers and missing values<sup>3-5</sup>. In addition, due to the temporal instability of parameters in traditional statistical models, the capabilities of these models in predicting crash occurrence and crash severity are limited. To overcome this limitation, the traditional statistical models have been extended to endogeneity and heterogeneity models. However, dimensionality challenges have been observed for large datasets when endogeneity and heterogeneity models are used<sup>6,7</sup>. Data-driven methods include a variety of techniques, among them Machine Learning (ML) techniques, which have progressed drastically over the past years. ML methods have few or no pre-assumed relationship between covariates, and they can better handle outliers as well

<sup>1</sup>Faculty of Civil & Environmental Engineering, Tarbiat Modares University, Tehran, Iran. <sup>2</sup>Civil and Infrastructure Engineering Discipline, RMIT University, Melbourne, Australia. ✉email: Seyedabrishami@modares.ac.ir

as the missing and noisy data<sup>8,9</sup>. There is a growing body of research on the analysis of road crash severity using ML techniques. In these studies, ML techniques are mainly used for the prediction of crash severity, including artificial neural networks<sup>10–14</sup>, tree-based models<sup>4,5,8,15–25</sup>, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), clustering methods, Naïve Bayes (NB), and hybrid models<sup>3,20,26–30</sup>. Heuristic and metaheuristic methods are also used in a few studies, and their results have been compared with machine learning-based models<sup>31,32</sup>.

Crash severity data are usually disproportionately distributed. The response variable is highly varied and imbalanced towards Property Damage Only (PDO) and less severe crashes. Traditional classification machine learning algorithms have a bias toward majority classes. However, due to the importance and the higher costs of severe and fatal crashes, there will be a significant loss in case of misclassification of severe and fatal crashes<sup>33</sup>. To address this problem, there are two solutions, namely data-level and algorithm-level solutions. At data-level solutions, usually, different sampling methods are used to create a balance between different classes of data. At algorithm-level methods, an ensemble methodology is used to improve the performance of a single classifier. In this study, the latter has recruited to deal with imbalanced and high variance of the used crash data.

The traffic profile data used in traditional crash severity models as independent variables are aggregate and are mainly based on static traffic data. These types of data are unable to explain variations in traffic profiles prior to the crash occurrence<sup>34</sup>. By emerging high-resolution traffic data, it is possible to capture detailed individual drivers' behavioural information and other influencers in traffic crashes<sup>35</sup>. By analysing microscopic driving behaviour through high-resolution data, it is possible to investigate movement variations before the crash<sup>36</sup>.

To enhance model robustness and improve the overall performance of the system, different ensemble methods have been recruited in the context of supervised learning. Dietterich<sup>37</sup> introduced ensemble methods as learning algorithms that develop a set of classifiers/regressors to classify/regress new data points by taking a vote of their predictions. In ensemble learning, a set of base learners are developed, and then, by combining them, a composite predictor is created<sup>38,39</sup>. Ensemble learning is also widely used in crash severity analysis. Geurts, and Loupe<sup>40</sup> developed a vote sequential minimal optimisation algorithm (an improved training algorithm based on SVMs) with bagging decision trees to analyse fatality occurrence in Lebanese road crashes, which outperformed sequential minimal optimisation, Naïve Bayes, logistic regression, artificial neural networks, Random Forest (RF), and bagging decision trees. Effati et al.<sup>20</sup> compared four models to predict motor vehicle crashes and their severity on two-lane, two-way roads, including Classification And Regression Tree (CART), Fuzzy Classification And Regression Tree (FCART), bagged-FCART, and SVM. They found that bagged-FCART provides higher overall accuracy and Kappa in comparison to the other three models. RF technique is one of the widely used ensemble learning methods in crash severity analysis. Mafi et al.<sup>4</sup> and Zhang et al.<sup>8</sup> used RF models in comparison to ordered probit, multinomial logit, K-nearest neighbour, decision tree, and SVM to predict the crash severities in Florida, U.S. Tang et al.<sup>9</sup> proposed a newly introduced two-layer stacking framework to analyse crash severity with 3-year crash data from 2004 to 2006 of Florida, U.S. In the first layer, three classifiers, including random forests, AdaBoost with decision tree base estimator, and Gradient Boosting Decision Tree (GBDT), were trained and then based on the outputs of the first layer, logistic regression was fitted as meta classifier in the second layer. Results showed that the stacking model had more accurate predictions compared to SVM, multi-layer perceptron and RF. In random forests, by randomly associating variables to each tree, the variance of the results will be reduced. However, in tree-based models, there are significant error rates originating from cut-point (the splitting threshold for each tree at each node) variance<sup>41</sup>. In random forests, the cut points are selected to optimise the quality of the split (i.e. Gini impurity or entropy), and therefore it cannot capture the cut-point variance. To overcome this limitation, Extremely Randomised Trees (ERTs) have been introduced by<sup>41</sup>. In contrast to RFs that use bootstrap sampling (i.e. random sampling with replacement), ERT uses the whole sample. Multiple studies in different fields like computer science, electrical engineering, chemical engineering, and energy systems have shown that ERT often can be considered a competitive alternative to RFs in terms of accuracy and other goodness of fit measures<sup>42–46</sup>. Crash data inherently contains different sources of variance. Variations in vehicle attributes, roadway characteristics, and environmental conditions affect both the likelihood and resulting injury severity of crashes. In addition, variance in drivers' physiologies and responses to the aforementioned sources can have an impact on crash severity<sup>47</sup>. With such highly variant data, ERT's ability to capture variances in data may result in better evaluation and prediction of the severity of crashes.

This paper contributes to the safety literature by providing an ERT for analysing inter-urban traffic crashes in Iran between 2013 and 2017. The crash data of rural roads in Khorasan Razavi, Iran, is used in this research. No previous research in the literature used ERT for analysing crash severity in rural areas. In this research, five types of variables, including traffic variables (e.g. average speed, traffic flow), vehicle types (e.g. motorcycle at fault), vehicle movement (e.g. moving forward, left turn), land use specifications (e.g. residential, agricultural), temporal characteristics (e.g. month, season), and environmental variables (e.g. road lighting, road and shoulder pavement condition, road surface condition) are used to develop the model.

The rest of this paper is structured as follows. The case study, including the data description, is presented in the following section. It is followed by explaining the specifications of ML classifiers and presenting the model. After that, the results are presented, and the findings are discussed in “[Results and discussions](#)” section. The final section presents concluding remarks and highlights directions of future research.

## Dataset

In this study, the crash dataset of rural roads in Khorasan Razavi, Iran, is used. The data is obtained from the Road and Maintenance Transportation Organisation in Iran for a period of 5 years, from 2013 to 2017. The major part of the crash data is extracted from police reports. The police reports contain information about injury severity, road geometry (e.g. shoulder type), weather condition, environmental condition (e.g. road lighting), crash cause, and vehicles and individuals involved in a crash. Real-time traffic data collected by inductive loop

detectors are also used in this research. The mean and standard deviation of loop detectors' speed, headway, and hourly traffic volumes are used in this research to represent real-time traffic situations before the occurrence of a crash. In this research, the crash severity is categorised into three ordered levels, including PDO, injury, and fatal. These levels are shown in Table 1.

Level 1 shows the property damage only crashes, accounting for 37% of observations. Level 2 includes all kinds of injuries that occur at the crash scene. Due to the wide variation of injury (from anticipating injury to possible injury), this level is the largest component of the data (59%). Level 3 belongs to fatal crashes, which has the lowest percentage (around 4%). 31 variables, including 28 binary variables and 3 numeric variables, are selected for modelling. A detailed description of these variables is shown in Table 2.

## Model development and interpretation

In this study, ERTs investigate the relationship between crash severity and the associated influencing factors. Similar to the RF model, the ERT is an ensemble technique in which a large number of decision trees are developed, and the most accurate tree is selected by the majority vote. In the following subsections, the ERT technique is explained.

**Extremely Randomised Trees (ERTs).** The ERT is a tree-based ensemble technique. In this technique, a set of classification trees are developed. In this study, Classification And Regression Trees (CARTs) are used as the base decision tree algorithm. In comparison to other decision tree algorithms like Iterative Dichotomiser 3 (ID3) and C4.5, CART can deal with both numerical and categorical variables, identify the most significant variables and eliminate insignificant ones, and handle outliers<sup>48</sup>. In this study, there are both numerical and categorical variables, and due to a large number of input variables, some of the insignificant variables should be discarded. In ERT, at each node, a random subset of  $K$  variables is selected from all input variables to decide on the split. Every tree is developed from the complete learning sample. Using the complete sample instead of bootstrap copying and random selection of cut-points are two key distinctive properties of ERT. The ERT technique has a number of parameters that should be tuned, among which three of them are more important. Those three parameters include: (1) the number of attributes that determine the strength of the future selection process (denoted by  $K$ ), (2) the minimum sample size for splitting a node that controls the strength of averaging output noise (shown by  $n_{\min}$ ), and (3) the number of trees (represented by  $M$ ) that decides variance reduction of the ensemble model<sup>41,49,50</sup>. For prediction, ERT similar to other ensemble methods, uses the combination of prediction results from each base model to predict the results for unseen data. In case of classification problems each base model contributes a vote for a category, and the category with the most votes wins for a given record in the list (majority rule)<sup>51,52</sup>.

**Tuning of parameters.** The cross-validation method is used to find the optimum values for the above mentioned three parameters. In addition to those 3 parameters, Minimum impurity Decrease (MiDe) is tuned. MiDe controls the depth of trees while fitting decision trees. The most accurate performing combination of these parameters is found through a grid search between all possible combinations. 5 values of  $M$ , 10 values of  $n_{\min}$ , two values of  $K$  ( $\sqrt{p}$  recommended by<sup>42</sup>,  $\log_2 p$ ) and 10 values of MiDe result in 1000 possible combinations of parameters. These combinations are examined in Python 3.6 via Scikit learn, Pandas, Numpy, and Matplotlib libraries.

**Model assessment.** In this study, the dataset is divided into train and test data to develop the ERTs. 80% of the dataset is randomly selected for training, and the remaining dataset is selected for testing. To evaluate the accuracy of results, four well-known measures extracted from the confusion matrix are used, including (1) Accuracy, (2) Precision, (3) Recall, and (4) F-measure. These measures are presented as follows.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False negative}} \quad (3)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Accuracy (Eq. 1) is the most common measure for classification problems<sup>53,54</sup>. The quality of the model is evaluated by accuracy through the proportion of all correct classifications over all observations. Accuracy is easily computable, understandable, and applicable for multi-class problems. However, this metric has some limitations. Due to the less distinctive value, it shows poor discriminating power in selecting an optimal classifier between a subset of models. In addition, it is inclined to majority class instances. Precision (Eq. 2) is the fraction of the valid classification of one class among predicted instances of that class. Recall (Eq. 3) is the ratio of retrieved and relevant observations to all relevant observations. Precision gives information on how well the model performs with regard to the False Positives. Recall, similar to precision, provides information about the performance with

Severity level		Train	Test	Total
Level 1	PDO-no injury	424 (37%)	104 (36%)	528 (37%)
Level 2	Injury	674 (59%)	166 (58%)	840 (59%)
Level 3	Fatal injury	43 (4%)	16 (6%)	59 (4%)
Overall		1141	286	1427

**Table 1.** Injury severity categories in the crash dataset.

Category	Variable	Type	Description	Train	Test
Vehicle	Moving forward	Binary	Crash occurred while moving forward		
			Zero	508 (45%)	120 (42%)
			One	633 (55%)	166 (58%)
	Motorcycle at fault	Binary	Motorcycle hit another vehicle		
			Zero	1034 (91%)	260 (91%)
			One	107 (9%)	26 (9%)
	Car involved in crash	Binary	At least one passenger car involved in crash		
			Zero	817 (72%)	212 (74%)
			One	324 (28%)	74 (26%)
	Hit fixed objects	Binary	Vehicle hit fixed objects		
			Zero	1044 (91%)	265 (93%)
			One	97 (9%)	21 (7%)
	Left turn	Binary	Crash occurred when one vehicle turning left		
			Zero	1079 (95%)	271 (95%)
			One	62 (5%)	15 (5%)
	Motorcycle innocent	Binary	Motorcycle was hit by another vehicle		
			Zero	1022 (90%)	254 (89%)
			One	119 (10%)	32 (11%)
	Truck at fault	Binary	Truck was found at fault in crash		
			Zero	1037 (91%)	260 (91%)
			One	104 (9%)	26 (9%)
	Truck innocent	Binary	Truck was not found at fault in crash		
			Zero	1082 (95%)	277 (97%)
			One	59 (5%)	9 (3%)
Pedestrian involved	Binary	Pedestrian was hit by vehicle			
		Zero	1057 (93%)	259 (91%)	
		One	84 (7%)	27 (9%)	
Land use	Residential area	Binary	Crash occurred in residential area		
			Zero	869 (76%)	219 (77%)
			One	272 (24%)	67 (23%)
	Agricultural area	Binary	Crash occurred in agricultural area		
Zero			1055 (92%)	268 (94%)	
One			86 (8%)	18 (6%)	
Crash cause	Fail to keep longitudinal distance	Binary	Fail to keep longitudinal distance		
			Zero	1077 (94%)	272 (95%)
			One	64 (6%)	14 (5%)
	Failing to yield the right of way	Binary	Driver failed to yield the right of way		
			Zero	1019 (89%)	260 (91%)
			One	122 (11%)	26 (9%)
Temporal	February	Binary	Crash occurred in February		
			Zero	1083 (95%)	262 (92%)
			One	58 (5%)	24 (8%)
	Winter	Binary	Crash occurred in winter		
			Zero	854 (75%)	199 (70%)
One	287 (25%)	87 (30%)			
Continued					

Category	Variable	Type	Description	Train	Test
Environment	No road marking	Binary	No marking condition was in the crash scene		
			Zero	937 (82%)	224 (78%)
			One	204 (18%)	62 (22%)
	Dry road surface	Binary	Dry road surface		
			Zero	323 (28%)	90 (31%)
			One	818 (72%)	196 (69%)
	Unpaved shoulder	Binary	Road had unpaved shoulder at crash scene		
			Zero	890 (78%)	203 (71%)
			One	251 (22%)	83 (29%)
	Clean weather (no cloud)	Binary	No cloud was in the sky		
			Zero	228 (20%)	73 (26%)
			One	913 (80%)	213 (74%)
	Road lighting deficiency	Binary	Lighting system was not fully functional		
			Zero	1133 (99%)	283 (99%)
			One	8 (1%)	3 (1%)
	Collision in the median strip	Binary	Collision occurred in the median strip		
			Zero	1052 (92%)	266 (93%)
			One	89 (8%)	20 (7%)
	Broken line	Binary	Broken marking line		
			Zero	402 (35%)	123 (43%)
			One	739 (65%)	163 (57%)
Direct-downhill road	Binary	Crash occurred in a direct-downhill road			
		Zero	1106 (97%)	278 (97%)	
		One	35 (3%)	8 (3%)	
Night without enough lighting	Binary	Crash occurred at night without enough lighting			
		Zero	892 (78%)	223 (78%)	
		One	249 (22%)	63 (22%)	
Snowy and frozen surface	Binary	Snowy and frozen road surface			
		Zero	1119 (98%)	277 (97%)	
		One	22 (2%)	9 (3%)	
Paved shoulder	Binary	The road had paved shoulder at crash scene			
		Zero	581 (51%)	165 (58%)	
		One	560 (49%)	121 (42%)	
Traffic	Flow rate (vehicle/min)	Continues	Traffic flow rate one hour before crash		
			Mean	8.07	8.1
			Standard deviation	8.34	8.44
	Headway (sec)	Continues	Headway one hour before crash		
			Mean	37.12	42.42
			Standard deviation	183.66	237.63
Average speed (km/h)	Continues	Average speed one hour before crash			
		Mean	80.05	79.53	
		Standard deviation	15.28	14.82	

**Table 2.** Description of the explanatory variables.

regards to the False Negatives. F-measure (Eq. 4), which is the harmonic mean, combines Precision and Recall. In comparison to accuracy, F-measure has a more distinctive power.

**Model-agnostic methods for model interpretation.** *Feature importance.* In tree-based models, the decrease in node impurity is weighted according to the ratio of the number of samples that reach a node to the total number of samples. This ratio is considered as feature importance. In a single decision tree model, the importance of a node is computed as follows.

$$node\_imp_i = w_i c_i - w_{l(i)} c_{l(i)} - w_{r(i)} c_{r(i)} \quad (5)$$

where:  $node\_imp_i$ : importance of node  $i$ ,  $w_i$ : weighted number of samples reaching node  $i$ ,  $c_i$ : the impurity value of node  $i$ ,  $l(i)$ : child node from left split on node  $i$ ,  $r(i)$ : child node from right split on node  $i$ .

Then importance of a feature in a decision tree is calculated using Eq. (6).

$$feat\_imp_j = \frac{\sum_{i:node\ i\ splits\ on\ feature\ j} node\_imp_i}{\sum_{k \in I} node\_imp_k} \quad (6)$$

where:  $feat\_imp_j$ : importance of feature  $j$ ,  $node\_imp_i$ : importance of node  $i$ ,  $I$ : set of nodes.

Normalised values of feature importance can be derived using the sum of all feature importance values.

**Partial dependence plots (PDPs).** Friedman<sup>55</sup> introduced PDPs in the context of gradient boosting machines. PDPs depict the behaviour of the dependent variable as a function of the selected independent variable(s). This is similar to the interpretation of coefficients from multiple linear regression. However, using PDPs make it possible to capture more complex relationships instead of a fixed coefficient<sup>56</sup>. To visualise the dependence of response variable and input variables and due to difficulties at higher dimensional arguments, partial dependence of response variable is defined as follows.

$$f_S(X_S) = E_{X_C} f(X_S, X_C) \quad (7)$$

where:  $f_S$ : marginal average of  $f$  over the distribution of feature  $X_C$ ;  $X_S$ : Subvector of variables of interest;  $X_C$ : Subvector of variables in which  $C$  is the complement set for  $S$ ,  $S \cup C = \{1, 2, \dots, p\}$  and  $p$  represents number of input variables.

Usually,  $X_S$  contains only one or two features and remaining appear in  $X_C$ . By averaging over all observations in train data partial dependence function can be estimated as:

$$\bar{f}_S(X_S) = \frac{1}{N} \sum_{i=1}^N f(X_S, x_{iC}) \quad (8)$$

where,  $\{x_{1C}, x_{2C}, \dots, x_{NC}\}$  are the values of  $X_C$  in train data<sup>38,57</sup>.

**Individual conditional expectation.** To disaggregate PDPs into individual levels, Goldstein et al.<sup>58</sup> introduced Individual Conditional Expectation (ICE) plots which display the estimated functional relationship for each observation. This individual-level estimation can capture the potential individual heterogeneity. Instead of drawing a global plot for all observations (in PDPs), ICE presents the visual relationship between the dependent variable and the variable of interest, which results in multiple lines per variable. The values for each line are computed by fixing all variables and drawing new values for the variable of interest from a grid and making predictions based on these new values<sup>57</sup>. The conditional expectation curve is defined as follows.

$$\bar{f}_S^i(X_S) = f(X_S, x_{iC}) \quad (9)$$

Since every ICE plot has its own intercept, it is difficult to distinguish heterogeneity in the model. To solve this problem, Goldstein et al.<sup>58</sup> proposed centred-ICE plots. In centred-ICE plots, a location  $x^*$  in the range of  $X_S$  is chosen and all predicted lines pinched to that point. It is shown that extremum values for  $x^*$  gives the most interpretable plots. For each  $\bar{f}_S^i$ , the corresponding centred-ICE plots curve is given by Eq. (10).

$$\bar{f}_{centered}^i = \bar{f}_S^i - 1(x^*, X_{iC}) \quad (10)$$

where:  $\bar{f}$ : fitted model,  $1$ : vector of 1's.

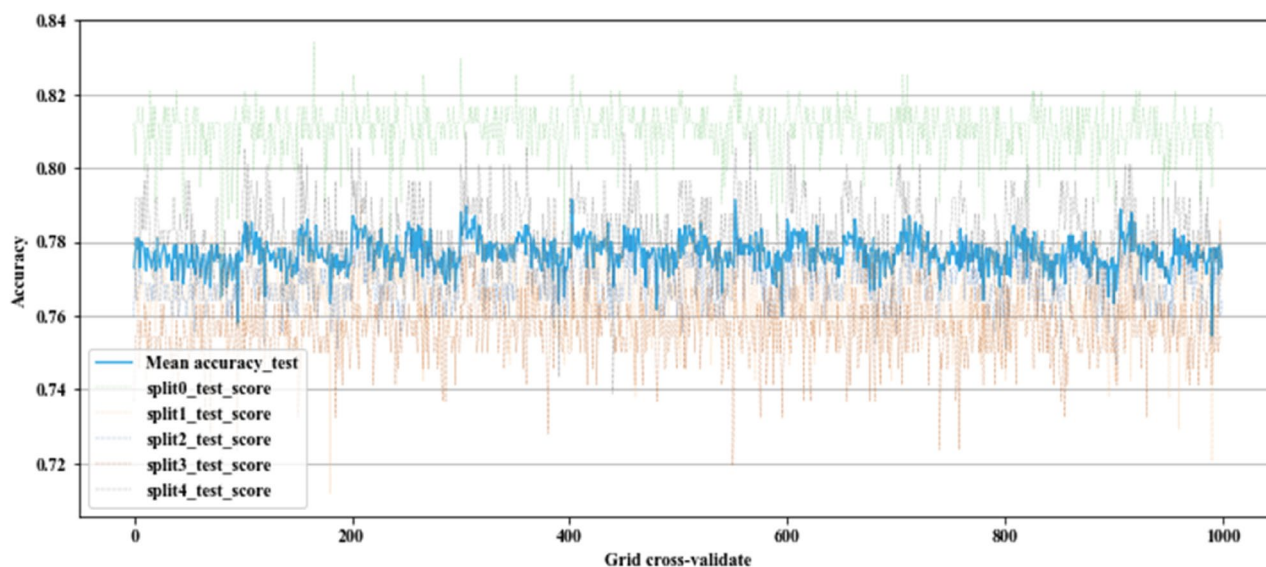
If  $x^*$  is the minimum values of  $X_S$ , it is enough for curve to originate at 0.

## Results and discussions

**ERT model.** In the ERT model development, 80% of the dataset is randomly selected as the train data to train the model, and the remaining 20% of the data is used as test data to examine the performance of the developed model. A cross-validation procedure is used to find the optimal parameters. In this procedure, the training dataset is divided into 5 smaller folds, and for every fold, a model is trained using fourfolds as training data and then the trained model is validated on the remaining fold. The average performance of this fivefold cross-validation is reported as model performance on the trained model. For each previously mentioned combination of parameters, a cross-validation procedure is executed. In Fig. 1, the results of the accuracy score for each combination of hyper-parameters are depicted. It can be seen that the model accuracy varies fold to fold and swings slightly between 0.72 and 0.84 (Fig. 1). The best set of parameters is achieved via  $\{M: 141, MiDe: 0.0, K: \sqrt{p}, nmin: 11\}$ .

To further evaluate the performance of the ERT model, a RF model is also developed, and the results from the RF model are compared with the results from the ERT model. Using a similar approach as the ERT model, fivefold cross-validation via the same set of hyperparameters has been recruited to calibrate the RF model. The comparison of the two models shows that the ERT model can outperform the other model, and the ERT model can be considered as a reliable competitor for RF models in crash severity analytics.

In Table 3, the final performance of the most accurate model is reported for test and train. By recruiting this set of parameters, the accuracy of 0.850 for the train and the accuracy of 0.814 for the test sets is achieved. Other measures also showed that there is little difference between the performance of the model on the train and test sets.



**Figure 1.** Accuracy score on different combination of hyper-parameters [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].

	Train			
	AdaBoost	XGBoost	Random forest	ERT
Confusion matrix	<i>PDO</i> $\begin{bmatrix} 291 & 133 & 0 \\ 98 & 575 & 0 \\ 8 & 35 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 282 & 142 & 0 \\ 70 & 604 & 0 \\ 6 & 37 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 310 & 114 & 0 \\ 38 & 636 & 0 \\ 5 & 35 & 3 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 326 & 98 & 0 \\ 36 & 638 & 0 \\ 5 & 30 & 8 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>
Accuracy	0.756	0.776	0.831	0.850
Precision	0.729	0.748	0.842	0.858
Recall	0.756	0.776	0.831	0.850
F-measure	0.743	0.758	0.816	0.840
	Test			
	AdaBoost	XGBoost	Random forest	ERT
Confusion matrix	<i>PDO</i> $\begin{bmatrix} 76 & 28 & 0 \\ 14 & 149 & 3 \\ 3 & 13 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 73 & 31 & 0 \\ 9 & 157 & 0 \\ 1 & 15 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 74 & 30 & 0 \\ 8 & 158 & 0 \\ 1 & 15 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>	<i>PDO</i> $\begin{bmatrix} 76 & 28 & 0 \\ 8 & 157 & 1 \\ 1 & 15 & 0 \end{bmatrix}$ <i>Injury</i> <i>Fatal</i>
Accuracy	0.787	0.805	0.811	0.815
Precision	0.752	0.769	0.775	0.781
Recall	0.787	0.804	0.811	0.815
F-measure	0.766	0.778	0.784	0.790

**Table 3.** Performance of selected models.

The comparison of confusion matrixes shows the ERT model generally outperforms the RF, XGBoost, and AdaBoost models. Focusing on less-frequent crash severity classes (e.g. fatal and PDO classes in this dataset), the ERT model slightly gained more accurate results than the other counterparts. Results from the ERT model show that this model, in comparison with the other models, is more inclined to the fatal class with the lowest frequency. The goodness of fit measures calculated in Table 3 indicates that the ERT model resulted in more accurate performance than its competitors even though the difference in measure values is low. These results show that the ERT model can be considered a competitive alternative for the RF, XGBoost, and AdaBoost models in classifying crash severity data.

Figure 2 shows Receiver Operating Characteristics (ROC) curves for each class, as well as the micro-average and macro-average curves. In Fig. 2, the nearest curve to random guess belongs to class 3 (fatal injury). However, two other classes (class 1 for PDO and class 2 for injury) gained higher true positive rates.

**Feature importance analysis.** The importance of variables used in model development is calculated using Eq. (6) and is shown in Fig. 3. According to the results, “moving forward” is the most important variable in the model. Moving forward as a variable indicating the direction affects collision type and consequently severity of the crash. This is consistent with the results from previous studies<sup>59</sup>. Vulnerable road users like motorcyclists and pedestrians are among the important variables in the model. Encountering vulnerable road users, especially

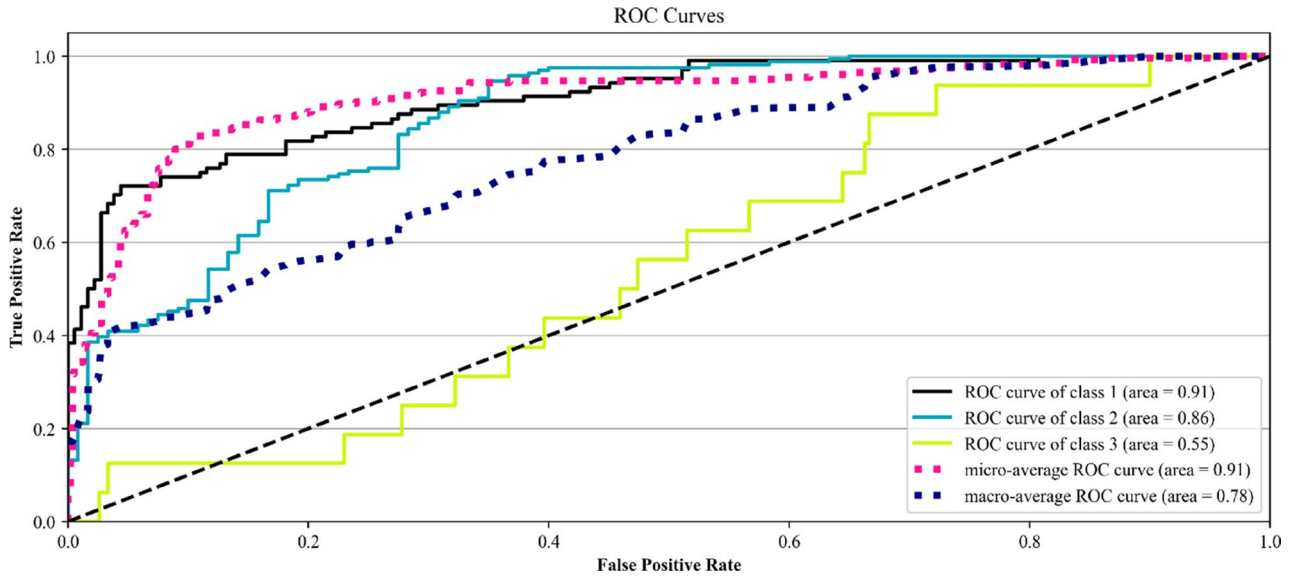


Figure 2. ROC curves for the ERT model [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].

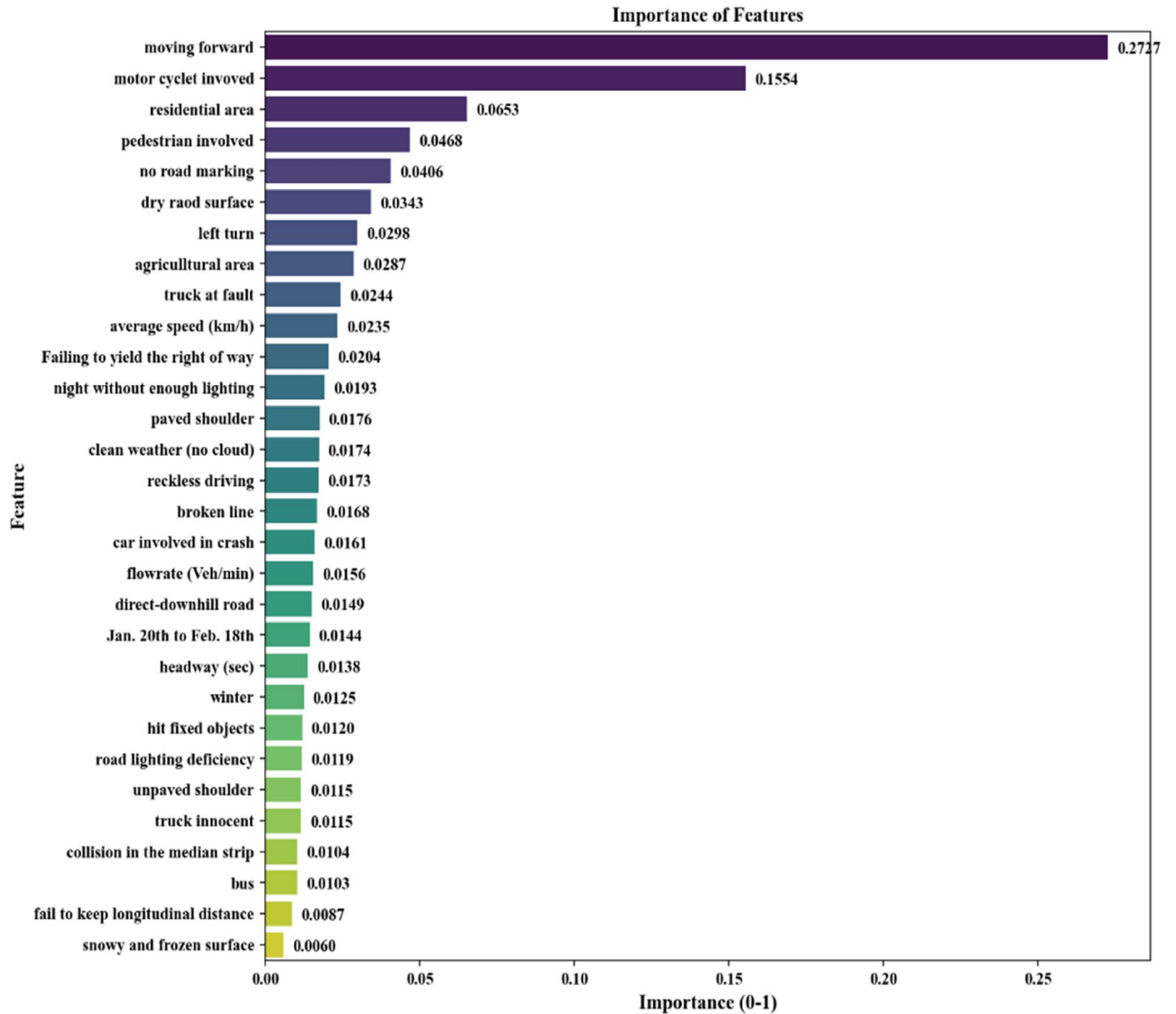


Figure 3. Importance of variables used in model [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].



in rural areas, can lead to more severe crashes. Similar outcomes resulted in previous studies<sup>60,61</sup>. Land use of crash location is another influencing variable on crash severity. Residential and agricultural areas are two main land use types in rural areas, which are also ranked 3rd and 8th in the list of important variables. In rural roads of Iran, the probability of facing distracted pedestrians and bicyclists and slowly moving agricultural machinery in such areas are higher. This is consistent with two other studies by<sup>16</sup> and<sup>162</sup>, in which Iranian and Malaysian crash data sets were recruited to develop decision trees and generalised ordered probit models to predict the crash severities.

One of the interesting results that can be derived from Fig. 3 is the importance of traffic-related variables. The average speed of traffic is found to be more important than 21 other variables. Limited studies have used traffic-related variables in crash severity analysis. In crash severity analysis conducted by<sup>28</sup> and<sup>63</sup>, traffic-related variables obtained higher importance.

Environmental variables are also among the influencing variables in crash severity in this research. For instance, no road marking and dry road surface are among the most important variables influencing crash severity in rural areas. However, the snowy and frozen surface is the least important variable in crash severity.

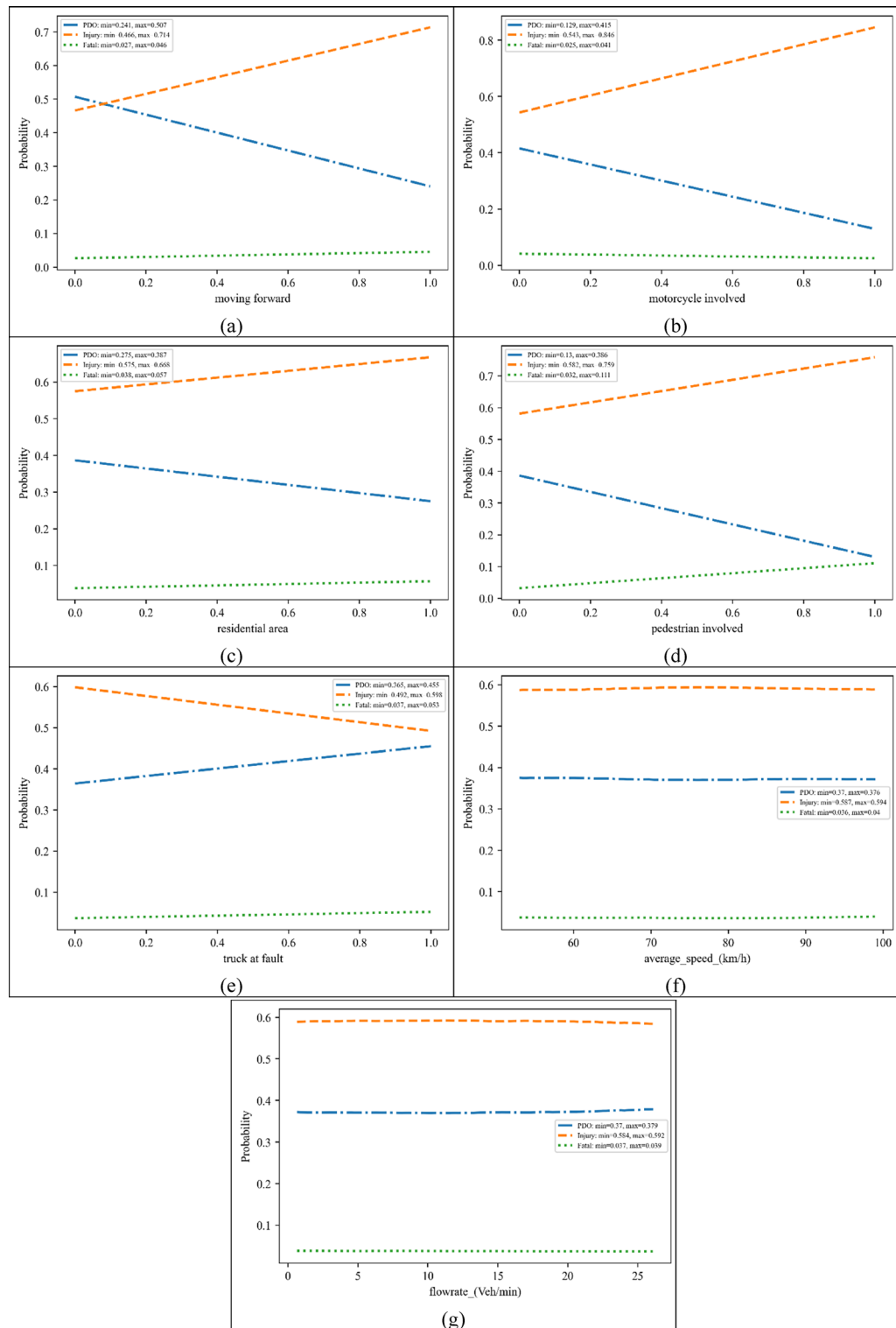
It can be seen that variables like road light deficiency and snowy surface with a very limited share in the dataset have lower values of feature importance, but these variables are very effective in identifying fatal crash categories and removing them from the data set significantly affects performance metrics of model. For instance, by removing road light deficiency and snowy surface, accuracy and precision in the train dataset would drop by 6 and 10%, respectively.

**Partial dependence plot (PDP).** PDPs for the top important variables are depicted in Fig. 4. From Fig. 4a, it can be seen that “moving forward” increases the probability of an injury crash. The proportion of fatal crashes is significantly low, but “moving forward” increases the probability of fatal crashes from 2.5 to 4.7%. Figure 4b shows that if a motorcycle gets involved in a crash, the probability of the occurrence of injury crashes is increased. Different safety strategies are applied in urban areas in Iran to improve the safety of motorcyclists. However, strategies need to be implemented to enhance the safety of motorcyclists in rural areas. In Fig. 4c, it can be seen that the existence of a residential section can raise the probability of a more severe crash. As mentioned in the previous section, in residential sections of rural areas, the chance of being faced with a distracted pedestrian or bicyclist is higher. It can be seen from Fig. 4d that if a pedestrian is hit by a vehicle, the probability of fatal crashes significantly increases. It emphasises that a series of appropriate countermeasures should be taken to address non-motorised safety on rural roads. A controversial result that can be concluded from Fig. 4e is the growth of PDO probability and reduction of the probability of injury crashes by the involvement of the “trucks at fault” variable. When a truck involves in a crash, there will be two possible collisions, including (1) a very severe collision that would end up in a fatal crash; (2) a non-significant collision in which no one is hurt. These trends are shown in Fig. 4e. With regard to traffic-related variables, it is noticeable that these variables have a slight impact on the probability of each crash category. The impact of average speed (Fig. 4f) on different categories of crashes is less than 1%. In traffic flows higher than 22 vehicles per minute (1,320 veh/h) (Fig. 4g), the probability of injury crashes reduces by almost 1%.

**Individual conditional expectation (ICE) plots.** ICE plots are developed for all observations in the train data set. Because of the large number of plots for each crash category (1141 curves), plots are depicted separately for each class in a single diagram. For the 4 most important variables, ICE plots are shown in Fig. 5. These plots show heterogeneity of crash severity. As observed in all plots in Fig. 5, all the curves start at various probabilities, and therefore the difference between observations is ambiguous. By centering plots around a certain point, individual heterogeneity can be uncovered more reliably.

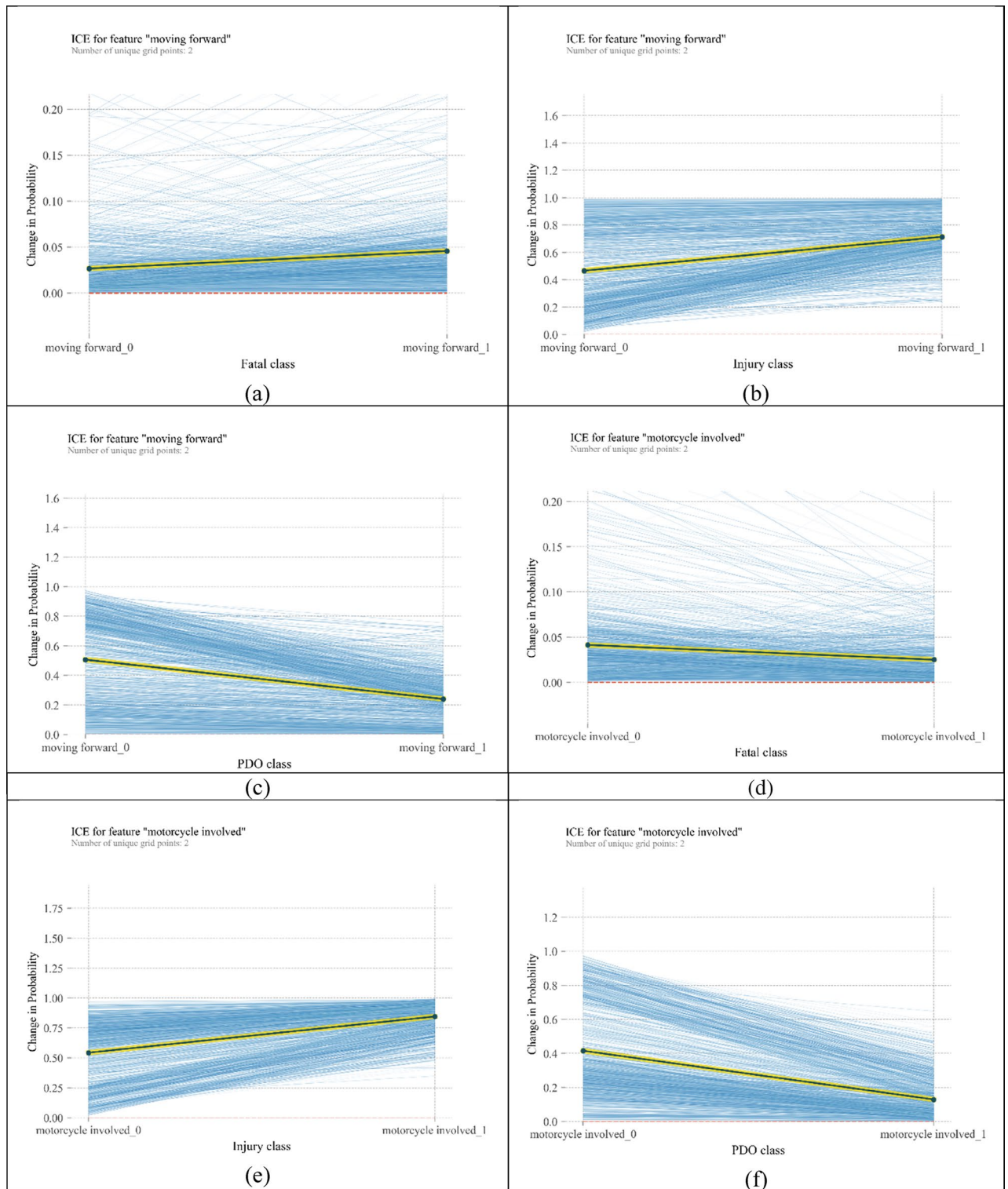
Centred ICE plots are shown in Fig. 6. Plots about “moving forward” show significant heterogeneity among observations, specifically in the fatal injury category. According to Fig. 6a, “moving forward” increases the probability of fatal crashes. From Fig. 6b, “moving forward” can double the probability of injury crashes; however, in some cases, this variable has no significant change in the probability of injury crashes. Figure 6c shows a heterogeneous but decreasing pattern of probability for the moving forward variable. It means that for all instances, getting involved in a crash while moving forward reduces its chance to be a PDO crash, but the amount of reduced probability differs. Similar heterogeneous behaviour can be seen for crashes involving motorcycles in Fig. 6d–f. However, in Fig. 6d, the average trend is decreasing. “Residential area” repeats the behaviour of the two above mentioned variables, but the magnitudes of changes in probabilities are bigger. In addition, in injury and fatal classes, there are some curves whose slope is against the slope of PDP. In other words, “residential area” as the location of the crash reflects more heterogeneity than “moving forward” and “motorcycle” variables. In relation to “pedestrian involved” variable, when a pedestrian is hit by a vehicle in some instances, the chance of fatality crashes increases by 25%, and the chance of injury crashes increases by 70%. However, in a few observations, the probability of injury crashes decreases. In the PDO class, the probability of crash occurrence can decline by 70%.

**Interaction of PDPs.** To uncover the interaction effect between independent variables, probability values by incorporating interaction effects are calculated. These values indicate how two variables could influence the dependent variable. Considering the interaction effects provide a better relationship between the dependent and independent variables<sup>64</sup>. Due to many possible plots for the combination of all variables for all categories of crashes ( $3 * \binom{30}{2} = 1305$ ), the interaction effects of the 4 most important variables and the average speed as the most important traffic-related variable are shown in Fig. 7.

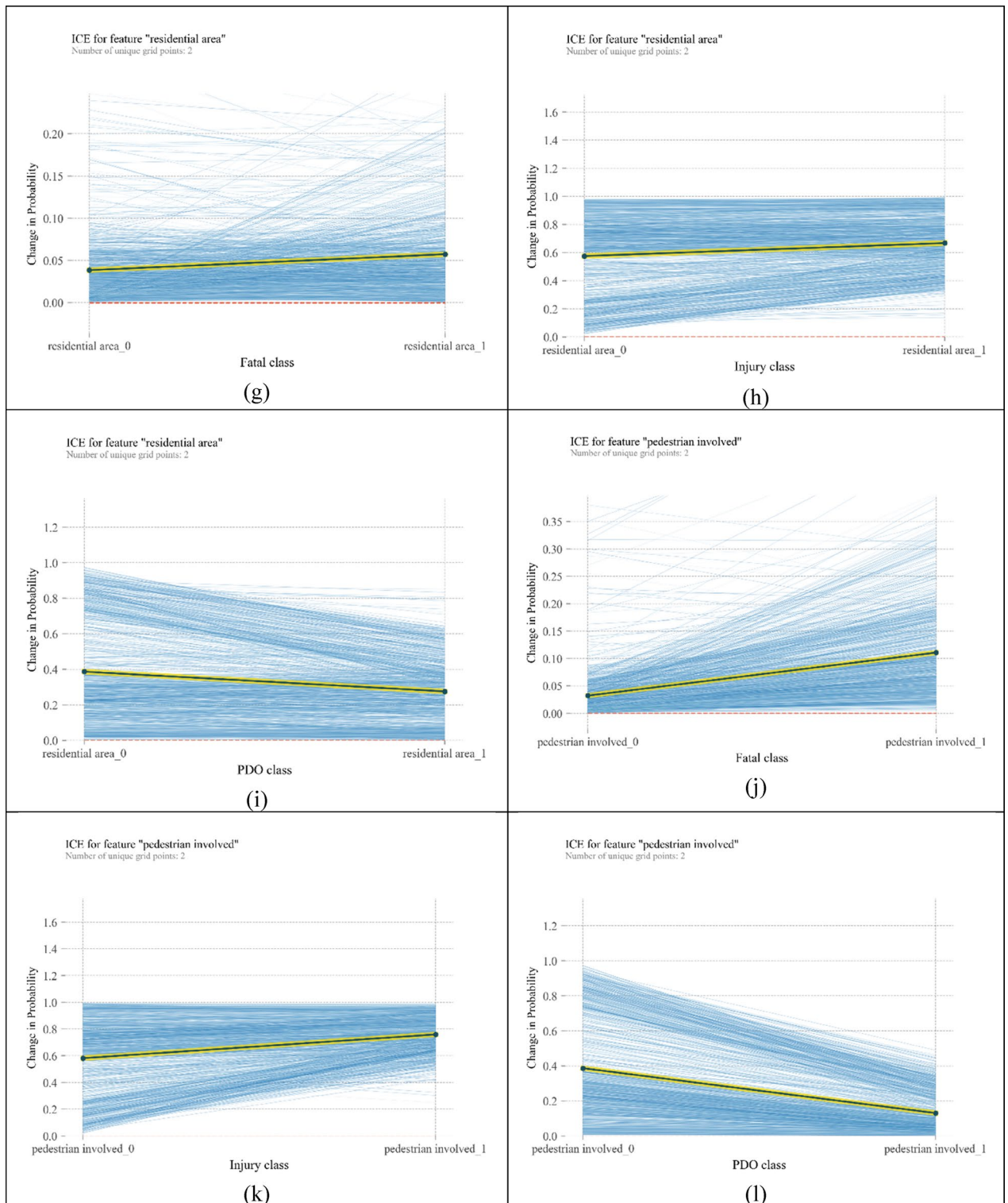


**Figure 4.** PDPs for the most influencing variables on crash severity [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].

In Fig. 7a, at lower speeds (from 20 to 40 km/h), by changing the moving direction to moving forward, there is no significant change in the probability of a fatal crash. However, at speeds of more than 90 km/h, the probability of fatal crashes is doubled. Almost parallel contours on the left side of Fig. 7b shows that changes in speed have a slight effect on the probability of injury crash, but its combination with moving forward could increase the probability of crash occurrence to 72% at speeds of 70 to 80 km/h. Smooth contours in Fig. 7c show that speed has no significant effect on the probability of a PDO crash. In Fig. 7d, the involvement of motorcycles in

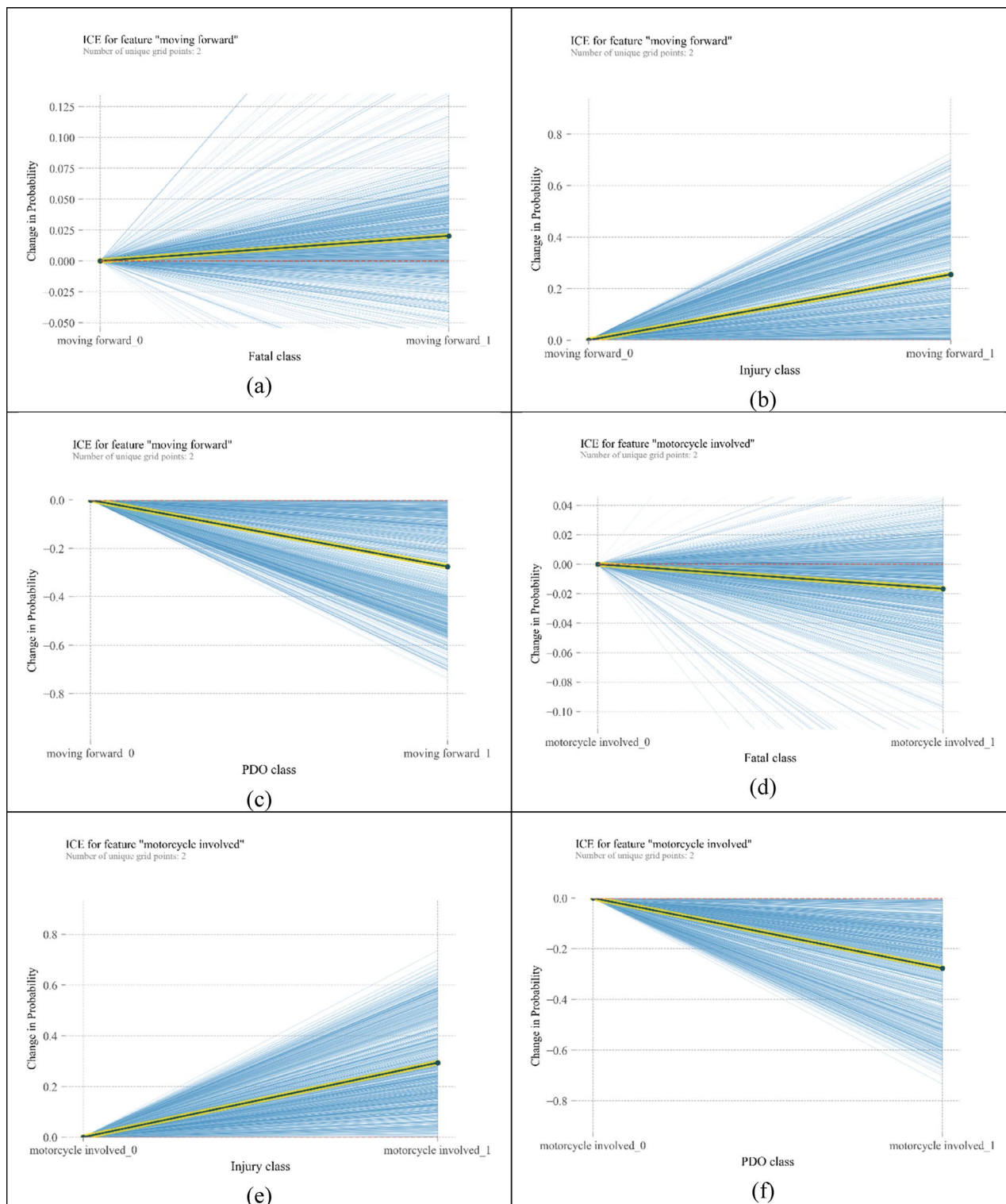


**Figure 5.** ICE plots for the most influencing variables on crash severity [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].

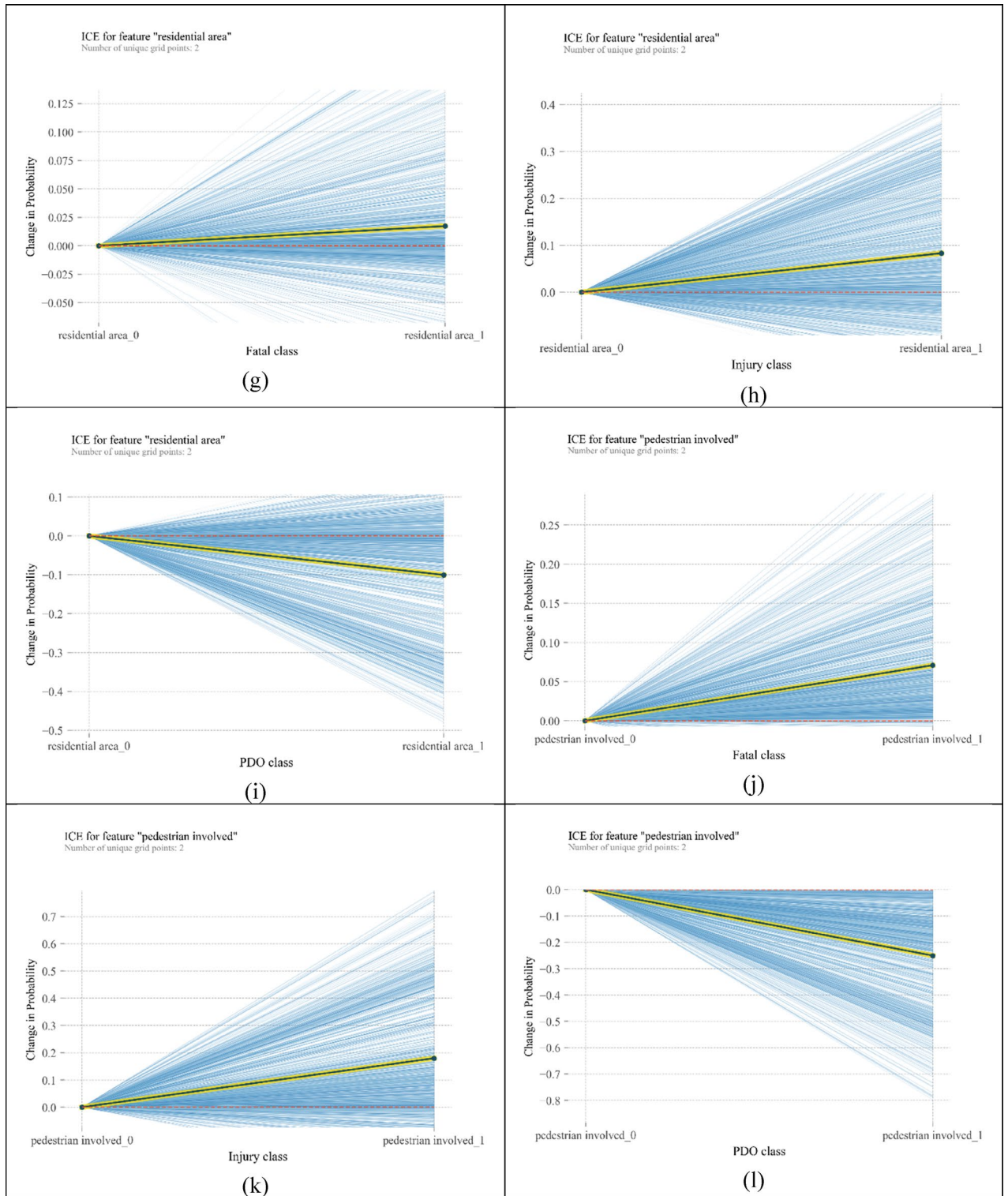


**Figure 5.** (continued)

crashes decreases the probability of fatal crashes. This can be due to a few fatal crash observations that contain motorcycles (4 fatal crash observations in 215 traffic crashes involving motorcycles). In motorcycle crashes and at speeds between 60 and 70 km/h, the probability of injury crash reaches its maximum as shown in Fig. 7e; however, it decreases at lower speeds. From Fig. 7f, it can be observed that a combination of motorcycles and higher speeds slumps the probability of a PDO crash, despite the minor effect of speed when there are no motorcycles involved in a crash. When a crash occurs in residential sections of a rural area, the probability of having a fatal crash is reduced with the increase in the average speed (Fig. 7g). However, in Fig. 7h opposite situation

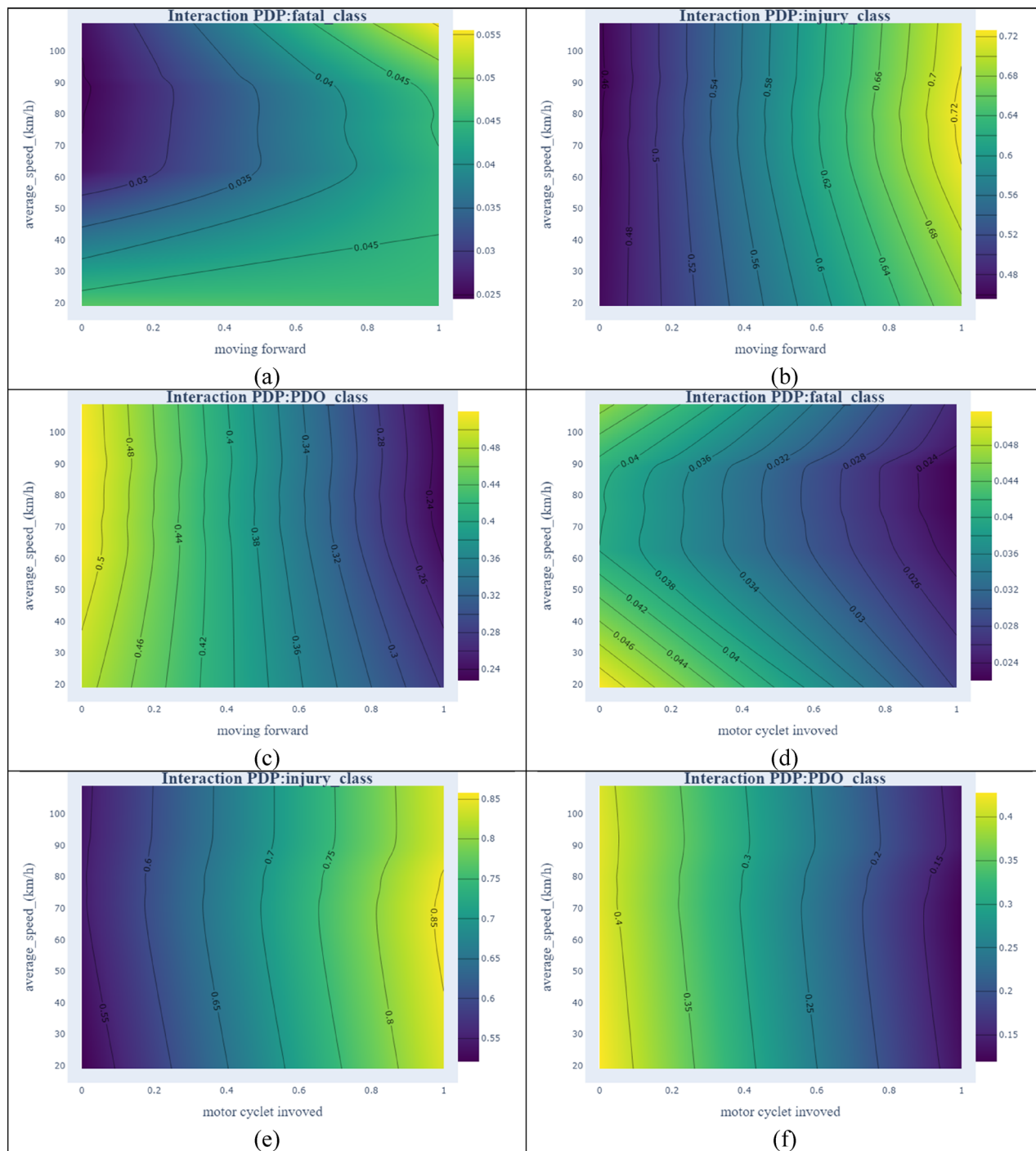


**Figure 6.** ICE plots for the most important variables [plotted by Python Matplotlib v. 3.3.4 <https://matplotlib.org>].

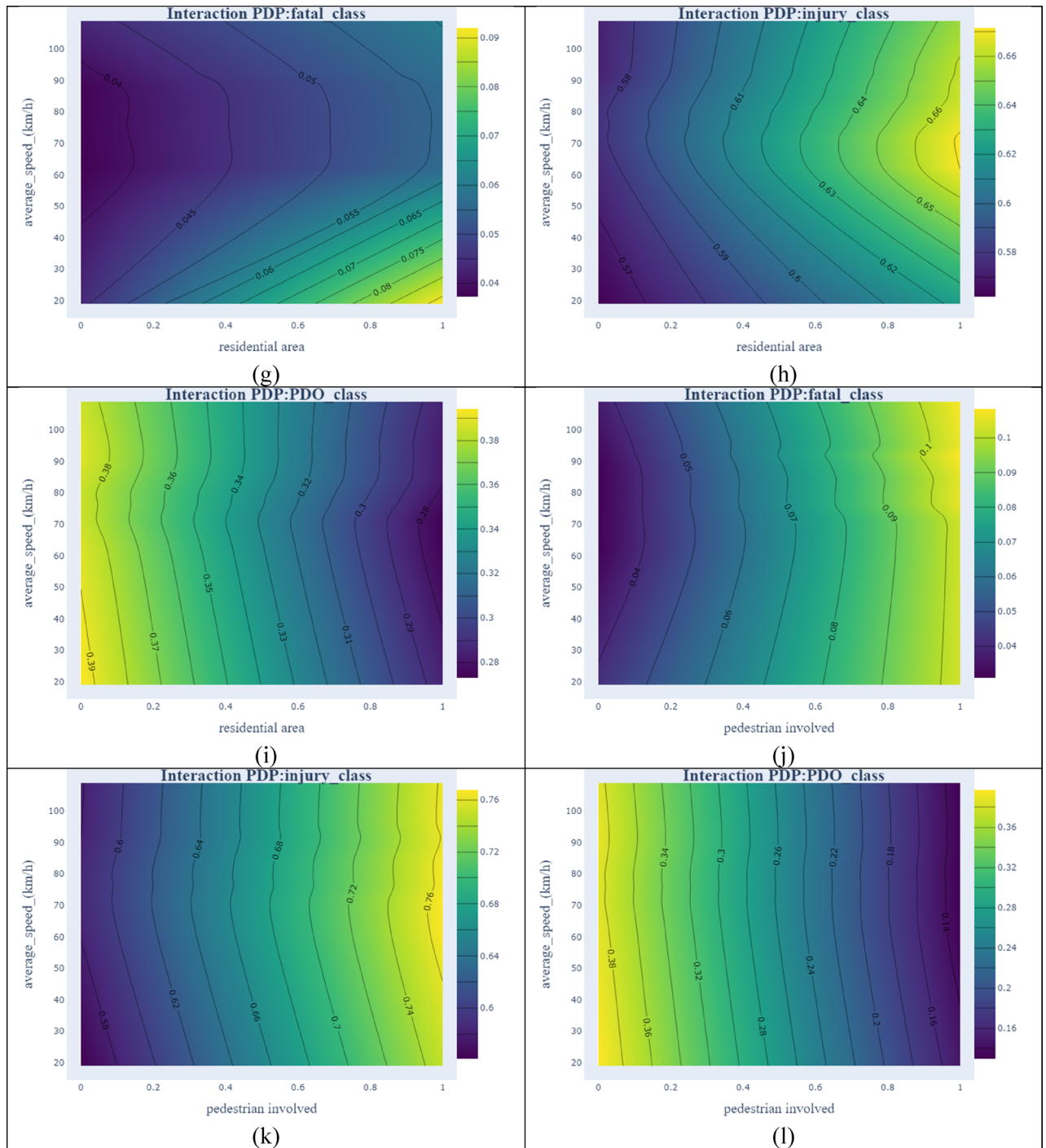


**Figure 6.** (continued)

is observed. High speed in residential areas can exacerbate the probability of injury crash to 0.66. In Fig. 7i, the probability of a PDO crash drops by the increase in speed and the existence of a residential area. However, vertical contours show that PDO crash probability is more dependent on the residential area than average speed. Figure 7j shows that fatal crash probability increases with pedestrian presence and speeding. At speeds around 70 km/h and when pedestrians are involved in the accident, the probability of injury crashes reaches the maximum value (Fig. 7k). The modest effect of speed on PDO crash probability can be derived from vertical contours in Fig. 7l.



**Figure 7.** Interaction PDPs of selected variables with average speed [plotted by Plotly].



**Figure 7.** (continued)

Across severity categories, it is noticeable that fatal crashes are more influenced by “average speed” compared to other categories of crash severity. However, average speed does not have a substantial impact on PDO crashes which is shown in Fig. 7 by vertical and almost parallel contours. With respect to injury crashes, regardless of the second variable in all figures, at speeds of around 70 to 80 km/h, the probability of observing injury crashes reaches its maximum.

### Conclusions and future research directions

In this study, ERTs were trained to model crash severity using historical crash data in the Khorasan-Razavi province in Iran. A total of 1141 observations were used to train the model, and 285 instances were recruited to test the model. The train and test models achieved an accuracy of 85% and 81% and the F-measure values of 0.840 and 0.790, respectively. Feature importance analysis showed that the direction of the movement alongside variables related to vulnerable users (e.g. pedestrians and motorcyclists) and land use variables (e.g. agricultural and



residential areas) were found as the most influencing variables on crash severity. In addition, the traffic variables (e.g. average speed, traffic flow, headway) have more influence on crash severity compared to other variables.

PDPs were depicted to show the changes in values of variables that have an impact on the average probability of each crash category. To uncover the heterogeneity across dataset, individual exception plots were structured for each instance. Interaction of partial dependences was plotted to investigate the combined effect of variables on the probability of each crash category. These plots provide a useful measure to interpret the machine learning models. Using these plots would enable researchers to dissect the nonlinear relationship between dependent and independent variables.

The performance of the ERT model showed that this model could be an alternative to the existing well-known models in traffic safety such as random forest, Adaboost, gradient boosting, and XGBoost. The involvement of motorcycles was found to be the most influential factor in PDO and injury crashes. In fatal crashes, the involvement of pedestrians had the highest impact on increasing the probability of a fatal crash. Policy implications can be derived from interpretation results. However, this study is one of the first applications of the ERT in crash safety analysis in rural areas, and it should be applied to different datasets. On the other hand, in this study, four parameters of the model were tuned. However, the ERT contains other parameters that have the potential to be investigated in future studies.

### Data availability

The sample dataset analysed during the current study are available in the GitHub repository at <https://github.com/farshidafshar/ERTModel>. The complete dataset is available from the corresponding author on reasonable request.

### Code availability

Code for data analysis is also available in the GitHub repository at <https://github.com/farshidafshar/ERTModel>.

Received: 2 November 2021; Accepted: 28 June 2022

Published online: 07 July 2022

### References

- Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study. IHME. (2018).
- World Bank. Guide for road safety opportunities and challenges: low and middle income country profiles. World Bank, Washington, DC. © World Bank. <https://openknowledge.worldbank.org/handle/10986/33363> License: CC BY 3.0 IGO (2020).
- Chen, C., Zhang, G., Qian, Z., Tarefder, R. A. & Tian, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accid. Anal. Prev.* **90**, 128–139 (2016).
- Mafi, S., Abdelrazig, Y. & Doczy, R. Machine learning methods to analyze injury severity of drivers from different age and gender groups. *Transp. Res. Rec.* **2672**(38), 171–183 (2018).
- Lamba, D., Alsadhan, M., Hsu, W., Fitzsimmons, E. & Newmark, G. Coping with class imbalance in classification of traffic crash severity based on sensor and road data: a feature selection and data augmentation approach. In *The 6th International Conference on Artificial Intelligence and Applications (AIAP-2019)*.
- Mannering, F., Bhat, C. R., Shankar, V. & Abdel-Aty, M. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Anal. Methods Accid. Res.* **25**, 100113 (2020).
- Bhowmik, T., Yasmin, S. & Eluru, N. A new econometric approach for modeling several count variables: A case study of crash frequency analysis by crash type and severity. *Transp. Res. Part B-Methodol.* **153**, 172–203 (2021).
- Zhang, J., Li, Z., Pu, Z. & Xu, C. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access.* **6**, 60079–60087 (2018).
- Tang, J., Liang, J., Han, C., Li, Z. & Huang, H. Crash injury severity analysis using a two-layer stacking framework. *Accid. Anal. Prev.* **122**, 226–238 (2019).
- Abdelwahab, H. T. & Abdel-Aty, M. A. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transp. Res. Rec.* **1746**(1), 6–13 (2001).
- Castro, Y. & Kim, Y. J. Data mining on road safety: factor assessment on vehicle accidents using classification models. *Int. J. Crash-worthiness* **21**(2), 104–111 (2016).
- Das, S., Dutta, A., Dixon, K., Minjares-Kyle, L. & Gillette, G. Using deep learning in severity analysis of at-fault motorcycle rider crashes. *Transp. Res. Rec.* **2672**(34), 122–134 (2018).
- Rezapour, M., Nazneen, S. & Ksaibati, K. Application of deep learning techniques in predicting motorcycle crash severity. *Eng. rep* **2**(7), 12175 (2020).
- Zheng, J. *et al.* Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access.* **7**, 39897–39910 (2019).
- Chong, M., Abraham, A. & Paprzycki, M. Traffic accident data mining using machine learning paradigms. In *Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04)*, 415–420 (2004).
- Kashani, A. T. & Mohaymany, A. S. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Saf. Sci.* **49**(10), 1314–1320 (2011).
- Krishnaveni, S. & Hemalatha, M. A perspective analysis of traffic accident using data mining techniques. *Int. J. Comput. Appl.* **23**(7), 40–48 (2011).
- Montella, A., Aria, M., D'Ambrosio, A. & Mauriello, F. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accid. Anal. Prev.* **49**, 58–72 (2012).
- Kashani, A. T., Rabiyeen, R. & Besharati, M. M. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. *J. Saf. Res.* **51**, 93–98 (2014).
- Effati, M., Thill, J. C. & Shabani, S. Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor. *J. Geogr. Syst.* **17**(2), 107–135 (2015).
- Toran Pour, A. T., Moridpour, S., Tay, R. & Rajabifard, A. Influencing factors on vehicle-pedestrian crash severity of school-aged pedestrian. *WIT Trans. Built Environ.* **176**, 465–475 (2017).
- Toran Pour, A., Moridpour, S., Tay, R. & Rajabifard, A. Modelling pedestrian crash severity at mid-blocks. *Transp. A Transp. Sci.* **13**(3), 273–297 (2017).
- Wahab, L. & Jiang, H. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS ONE* **14**(4), e0214966 (2019).

24. Rahman, M. S., Abdel-Aty, M., Hasan, S. & Cai, Q. Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. *J. Saf. Res.* **70**, 275–288 (2019).
25. Lin, C., Wu, D., Liu, H., Xia, X. & Bhattarai, N. Factor identification and prediction for teen driver crash severity using machine learning: A case study. *Appl. Sci.* **10**(5), 1675 (2020).
26. Li, Z., Liu, P., Wang, W. & Xu, C. Using support vector machine models for crash injury severity analysis. *Accid. Anal. Prev.* **45**, 478–486 (2012).
27. Kusano, K. & Gabler, H. C. Comparison and validation of injury risk classifiers for advanced automated crash notification systems. *Traffic Inj. Prev.* **15**(1), 126–133 (2014).
28. Yu, R. & Abdel-Aty, M. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Saf. Sci.* **63**, 50–56 (2014).
29. Ahmadi, A., Jahangiri, A., Berardi, V. & Machiani, S. G. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Transp. Saf. Secur.* **12**(4), 522–546 (2020).
30. Cheng, W. *et al.* Exploring the modeling and site-ranking performance of Bayesian spatiotemporal crash frequency models with mixture components. *Accid. Anal. Prev.* **135**, 105357 (2020).
31. Hasheminejad, S. H. A., Zahedi, M. & Hasheminejad, S. M. H. A hybrid clustering and classification approach for predicting crash injury severity on rural roads. *Int. J. Inj. Contr. Saf. Promot.* **25**(1), 85–101 (2021).
32. Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A. & Hadi, M. Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transp. Res. Rec.* **2673**(11), 680–692 (2019).
33. Peng, Y., Li, C., Wang, K., Gao, Z. & Yu, R. Examining imbalanced classification algorithms in predicting real-time traffic crash risk. *Accid. Anal. Prev.* **144**, 105610 (2020).
34. Xu, C., Wang, Y., Liu, P., Wang, W. & Bao, J. Quantitative risk assessment of freeway crash casualty using high-resolution traffic data. *Reliab. Eng. Syst.* **169**, 229–311 (2018).
35. Wu, X. & Liu, H. X. Using high-resolution event-based data for traffic modeling and control: An overview. *Transp. Res. Part C Emerg.* **42**, 28–43 (2014).
36. Arvin, R., Kamrani, M. & Khattak, A. J. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. *Accid. Anal. Prev.* **132**, 105226 (2019).
37. Dietterich, T.G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 1–15 (2000).
38. Hastie, T., Tibshirani, R. and Friedman, J. Additive models, trees, and related methods. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).
39. Zhu, Z., Wang, Z., Li, D., Zhu, Y. & Du, W. Geometric structural ensemble learning for imbalanced problems. *IEEE Trans. Cybern.* **50**(4), 1617–1629 (2018).
40. Ghandour, A. J., Hammoud, H. & Al-Hajj, S. Analyzing factors associated with fatal road crashes: A machine learning approach. *Int. J. Environ. Res.* **17**(11), 4111 (2020).
41. Geurts, P. & Louppe, G. Extremely randomized trees. *Mach. Learn.* **63**(1), 3–42 (2006).
42. Geurts, P. & Louppe, G. Learning to rank with extremely randomized trees. *Proc. Mach. Learn.* **14**, 49–61 (2011).
43. Dev, V. A., Datta, S., Chemmangattuvalappil, N. G. & Eden, M. R. Comparison of tree based ensemble machine learning methods for prediction of rate constant of Diels-Alder reaction. In *Computer Aided Chemical Engineering* (Elsevier, 2017).
44. Ahmad, M. W., Reynolds, J. & Rezgui, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *J. Clean. Prod.* **203**, 810–821 (2018).
45. Gu, Y. *et al.* Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Sci. Rep.* **11**(1), 1–13 (2021).
46. Piccialli, F. *et al.* Precision medicine and machine learning towards the prediction of the outcome of potential celiac disease. *Sci. Rep.* **11**(1), 1–10 (2021).
47. Mannering, F. L., Shankar, V. & Bhat, C. R. Unobserved heterogeneity and the statistical analysis of highway accident data. *Anal. Methods Accid. Res.* **11**, 1–16 (2016).
48. Singh, S. & Gupta, P. Comparative study ID3, cart and C4. 5 decision tree algorithm: A survey. *IJAIST.* **27**(27), 97–103 (2014).
49. Eslami, E., Salman, A. K., Choi, Y., Sayeed, A. & Lops, Y. A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. *Neural Comput. Appl.* **32**(11), 7563–7579 (2019).
50. Paul, A., Furmanchuk, A., Liao, W. K., Choudhary, A. & Agrawal, A. Property prediction of organic donor molecules for photo-voltaic applications using extremely randomized trees. *Mol. Inform.* **38**(11–12), 1900038 (2019).
51. Nisbet, R., Elder, J. & Miner, G. D. Model Evaluation and Enhancement. In *Handbook of Statistical Analysis and Data Mining Applications (second edition)*. (Academic Press, 2018).
52. Kotu, V. & Deshpande, B. Data mining process. In *Predictive Analytics and Data Mining: Concepts and Practice with Rapidminer* (Morgan Kaufmann, 2014).
53. Gu, Q., Zhu, L. & Cai, Z. Evaluation measures of the classification performance of imbalanced data sets. In *International Symposium on Intelligence Computation and Applications*, 461–471 (2009).
54. Hossin, M. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *IJDKP.* **5**(2), 1 (2015).
55. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
56. Biecek, P. & Burzykowski, T. Partial-dependence Profiles. In *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models* (CRC Press, 2021).
57. Molnar, C. Interpretable machine learning. A Guide for Making Black Box Models Explainable <https://christophm.github.io/interpretable-ml-book>. (2019).
58. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**(1), 44–65 (2015).
59. Jamal, A. & Umer, W. Exploring the injury severity risk factors in fatal crashes with neural network. *Int. J. Environ. Res.* **17**(20), 7446 (2020).
60. Savolainen, P. T., Mannering, F. L., Lord, D. & Quddus, M. A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accid. Anal. Prev.* **43**(5), 1666–1676 (2011).
61. Cantillo, V., Márquez, L. & Díaz, C. J. An exploratory analysis of factors associated with traffic crashes severity in Cartagena, Colombia. *Accid. Anal. Prev.* **146**, 105749 (2020).
62. Anarkooli, A. J., Hosseinpour, M. & Kardar, A. Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. *Accid. Anal. Prev.* **106**, 399–410 (2017).
63. Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S. & Mohammadian, A. K. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **136**, 105405 (2020).
64. Lavrakas, P. J. Survey statistics. In *Encyclopedia of Survey Research Methods* (Sage Publications, 2008).

## Acknowledgements

Thanks to the Road Management Centre of Iran and the Traffic Safety Office of Road and Maintenance Transportation Organization of Khorasan Razavi Province for providing traffic and crash data for research.

### Author contributions

F.A. wrote the main manuscripts and S.S. and S.M. supervise the process of data analysis and results' discussion and finally reviewed the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to S.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022