Accident Analysis and Prevention 160 (2021) 106322

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors

Zhengjing Ma [a], Gang Mei [a,*], Salvatore Cuomo [b]

[a] School of Engineering and Technology, China University of Geosciences (Beijing), Beijing 100083, China
[b] Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Italy

ARTICLE INFO

Keywords:
Road safety
Traffic accidents
Injury severity
Deep learning

ABSTRACT

Vulnerable road users (VRUs) are exposed to the highest risk in the road traffic environment. Analyzing contributing factors that affect injury severity facilitates injury severity prediction and further application in developing countermeasures to guarantee VRUs safety. Recently, machine learning approaches have been introduced, in which analyses tend to be one-sided and may ignore important information. To solve this problem, this paper proposes a comprehensive analytic framework that employs a deep learning model referred to as the stacked sparse autoencoder (SSAE) to predict the injury severity of traffic accidents based on contributing factors. The essential idea of the method is to integrate various analyses into an analytical framework that performs corresponding data processing and analysis by different machine learning approaches. In the proposed method, first, we utilize a machine learning approach (i.e., Catboost) to analyze the importance and dependence of the contributing factors to injury severity and remove low correlation factors; second, according to the geographical information, we classify the data into different classes by utilizing a machine learning approach (i.e., *k*-means clustering); third, by employing high correlation factors, we employ an SSAE-based deep learning model to perform injury severity prediction in each data class. By experiments with a real-world traffic accident dataset, we demonstrated the effectiveness and applicability of the framework. Specifically, (1) the importance and dependence of contributing factors were obtained by CatBoost and the Shapley value, and (2) the SSAE-based deep learning model achieved the best performance compared to other baseline models. The proposed analytic framework can also be utilized for other accident data for severity or other risk indicator analyses involving VRUs safety.

1. Introduction

Road safety represents a growing socioeconomic challenge worldwide. According to the *Global Status Report on Road Safety* published by the World Health Organization in 2018, road traffic accidents are responsible for approximately 1.35 million traffic fatalities and 20–50 million non-fatal injuries worldwide each year. The highest percentages of the total number of injuries and fatalities among the different groups of road users involved are pedestrians, cyclists, and motorcyclists. This group of road users that tend to be more vulnerable are commonly referred to as vulnerable road users (VRUs) (Vanlaar et al., 2016).

Given the varying levels of injuries involved in traffic accidents, a commonly evaluated indicator is injury severity, which is one of the critical indicators/measures used for assessing road safety performance (Riveiro et al., 2017). To alleviate the adverse consequences of traffic accidents for VRU safety, it is imperative to analyze traffic accident data to investigate the relationship between injury severity outcomes and their related risk factors and to develop additional predictive models for injury severity. Typically, injury severity is considered to be associated with a range of risk factors (e.g., driver characteristics, crash and roadway factors, vehicle characteristics, and environmental characteristics) (Wali et al., 2020; Behnood and Mannering, 2015; Ahmad et al., 2019). Knowledge of how these contributing factors lead to increased injury severity facilitates exploring injury patterns and adopting

Abbreviations: AE, Autoencoder; ANN, Artificial Neural Networks; CART, Classification And Regression Trees; CNN, Convolutional neural networks; DNN, Beep Neural Networks; DT, Decision Tree; GBDT, Gradient Boosted Decision Trees; LSTM, Long Short-Term Memory; RF, Random Forest; SHAP, Shapley additional explanation; SSAE, Stacked Sparse Autoencoder; SVM, Support Vector Machine; VRUs, vulnerable road users; XGBoost, eXtreme Gradient Boosting.

* Corresponding author.
E-mail address: gang.mei@cugb.edu.cn (G. Mei).

https://doi.org/10.1016/j.aap.2021.106322
Received 30 March 2021; Received in revised form 22 July 2021; Accepted 23 July 2021
Available online 5 August 2021
0001-4575/© 2021 Elsevier Ltd. All rights reserved.

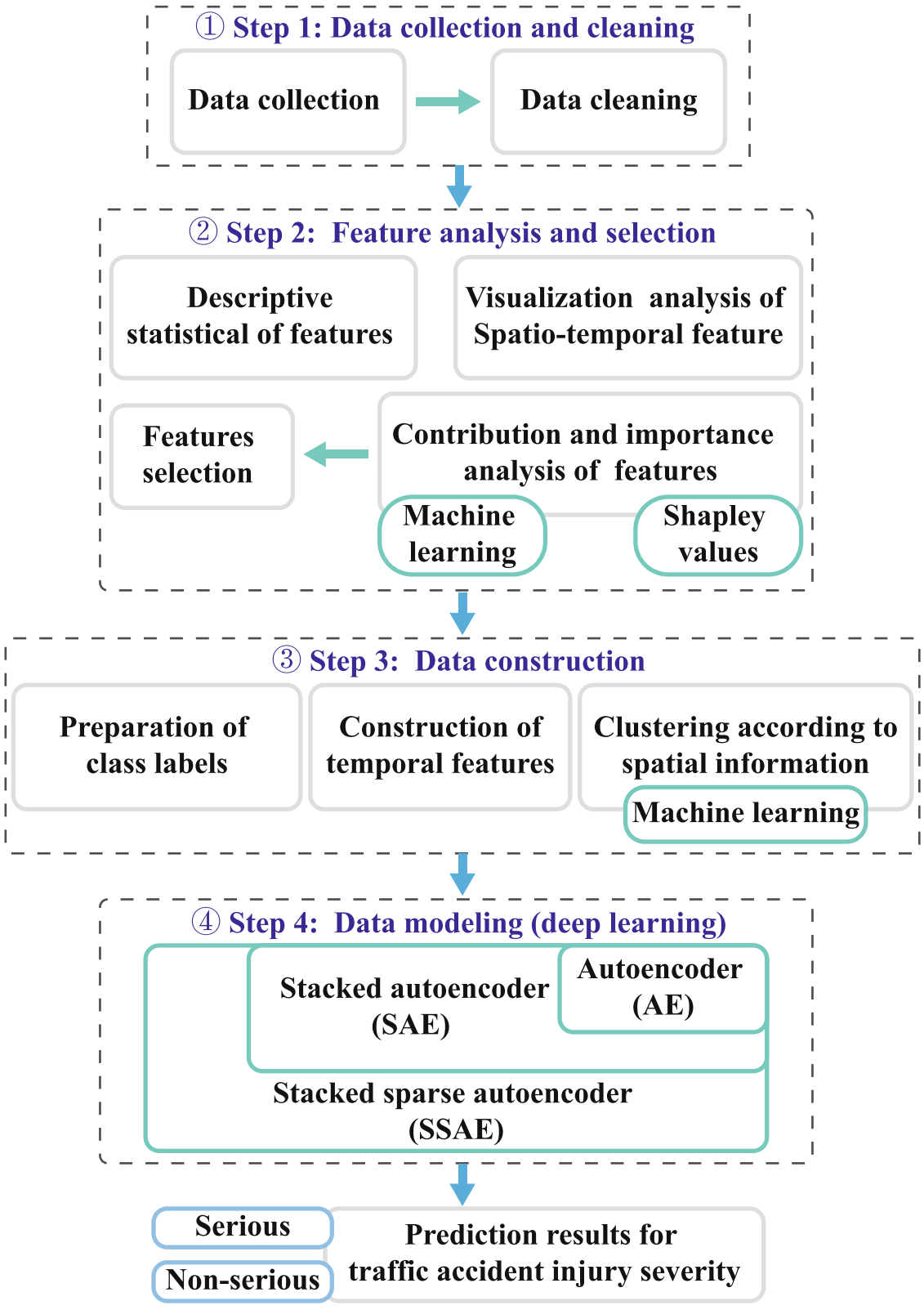

Fig. 1. An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors.

evidence-based safety improvement measures.

Some studies have been devoted to identifying and quantifying the impact of these contributing factors by common data analysis methods (Osama and Sayed, 2017; Yasmin et al., 2014; Eluru et al., 2008; Macpherson et al., 2004; Kim et al., 2007), for example, binary logit models (Yan et al., 2011) and the logit model (Xu et al., 2019). Recently, as an alternative to traditional data analysis methods, machine learning approaches have been gradually introduced to analyze contributing factors related to injury severity. It allows the extraction of valuable information from large quantities of complex and heterogeneous data. Common machine learning approaches that have been applied include decision trees (DT) (Taamneh et al., 2017), Bayesian networks (Ma et al., 2018;

Mujalli and De Oña, 2011), classification and regression trees (CART) (Dong and Zhou, 2020), random forest (RF) (Liu et al., 2020), and eXtreme gradient boosting (XGBoost) (Assi et al., 2020).

The analysis of accident data from different circumstances by machine learning approaches facilitates determining the importance of contributing factors of injury severity, which further facilitates selecting appropriate input data for developing predictive models. Predictive models refer to approaches that predict the injury severity in accidents by leveraging highly correlated factors as input features.

Injury severity is a complex phenomenon influenced by a variety of contributing factors, which means that the major challenge in developing predictive models is the nonlinear relationship between injury severity and multiple factors in an accident. Considering its proficiency in capturing the nonlinear relationship between input and output data, machine learning approaches have been employed as a predictive model for injury severity analysis, including DT (Abellán et al., 2013; De Oña et al., 2013a), Support Vector Machine (SVM) (Dong et al., 2015; Iranitalab and Khattak, 2017; Yu and Abdel-Aty, 2014), RF (Das et al., 2009), Bayesian networks (De Oña et al., 2013b), *k*-means clustering (Anderson, 2009; Mauro et al., 2013), and Artificial Neural Networks (ANN) (Zeng and Huang, 2014). Furthermore, as an emerging analytics technique, deep learning is increasingly being introduced into accident analysis, where commonly employed deep learning models include Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) (Jiang et al., 2020; Formosa et al., 2020; Huang et al., 2020; Li et al., 2020; Bichicchi et al., 2020; Bao et al., 2019).

Although these conventional machine learning approaches perform well, there are several limitations.

First, these methods have their respective emphasis in injury severity analysis, for instance, to analyze the impact of contributing factors or to predict the injury severity level. Second, as a black-box approach, most machine learning approaches deficiently explain the relationship between the contributing factors and injury severities. Third, the spatial correlation of these accidents' occurrence is commonly ignored. Fourth, models learning from datasets that include a small number of serious accidents are prone to bias: a higher predictive ability for non-serious injuries and a weaker predictive ability for serious accidents.

To address the above issues, this paper proposes an analytic framework that uses a deep learning model referred to as a stacked sparse autoencoder (SSAE) for predicting injury severity of traffic accidents. The framework's comprehensive analysis is based on contributing factors collected from accident records. The essential idea of the method is to integrate various analyses of traffic accident injury severity into a comprehensive analytic framework and perform the corresponding processing and analysis of accident data by different machine learning approaches.

The contributions of this paper can be summarized as follows.

(1) We employed a machine learning approach with interpretation (i.e., CatBoost combined with the Shapley value) for the analysis of contributing factors in traffic accident data, and revealed the contribution of different factors to injury severity in traffic accidents. The result is available to select highly correlated contributing factors for developing a predictive model.

(2) We considered the spatial correlation of accident locations and adopted a machine learning approach (i.e., *k*-means clustering) to classify accident data into four classes by utilizing geographic information to apply it in developing a predictive model separately.

(3) For each clustered result, we employed the SSAE-based deep learning model to predict the injury severity by applying constructed input. It allows us to reduce the feature space and further remove redundant and irrelevant information. Furthermore, it is able to handle the imbalanced data problem.

(4) We evaluated the proposed analytic framework by applying it to a UK accident record containing considerable spatial and temporal information and other contributing factors, and the result confirmed its effectiveness and applicability.

The rest of this paper is organized as follows. Section 2 describes the proposed method in detail. Section 3 applies the method in a real case and analyzes the results. Section 4 discusses the advantages and shortcomings of the proposed method, and the potential future work. Section 5 concludes the paper.

2. Methods

2.1. Overview

In this paper, we propose an analytic framework that employs a deep learning model referred to as SSAE for predicting injury severity of traffic accidents based on contributing factors. As demonstrated in Fig. 1, first, we collect a large quantity of traffic accident records from publicly available websites and clean these raw data. Second, we employ a machine learning approach (i.e., CatBoost) combined with the Shapley value to analyze the importance and dependence of the contributing factors for injury severity and remove low correlation factors. Third, we classify cleaned data into different classes by utilizing a machine learning approach (i.e., *k*-means clustering) based on the geographical information in traffic accident data. Fourth, based on the selected highly correlated contributing factors and the constructed temporal features, for each class of clustered results, we employ an SSAE-based deep learning model to predict injury severity of traffic accidents.

2.2. Step 1: Data collection and cleaning

In recent years, with rapid advances in monitoring system equipment (e.g., detectors and sensors), intelligent transportation systems (ITSs), and connected/automated vehicles (CAVs), the data available for traffic injury severity analysis have increased dramatically. Most of these datasets are available from public websites.

To obtain information from the collected data, it is important to clean the data. More specifically, first, the quality of the data should be checked for identifying incorrect, inconsistent, missing, and skewed information. Data profiling is a common method, which explores the quality of the data by summary statistics. The method can be employed to detect missing values, calculate correlations between variables and obtain statistics on the distribution of individual numerical variables, and thus identify missing values, outliers, and uncorrelated values (Azeroual et al., 2018). Second, the raw data requires to be cleaned up, including removing incorrect values and handling inconsistent or missing values.

2.3. Step 2: Feature analysis and selection

Features correspond to different variables in the collected and cleaned traffic accident data that influence the target. High-quality features have the advantages of being informative, relevant, interpretable, and nonredundant, which are the foundation for modeling and problem solving and for producing reliable and compelling results (Shi et al., 2019). Therefore, it is necessary to perform feature analysis and selection to obtain high-quality features for improved model performance.

Feature analysis involves summarizing and analyzing various contributing factors within the cleaned data, typically achieved through data exploration and visualization. More specifically, feature analysis mainly employs descriptive statistics and graphical techniques (e.g., plotting to obtain scatter plots, bar charts, and heat maps) (Kamruzzaman et al., 2019; Smith, 2016; Abdel-Aty et al., 2011). Furthermore, machine learning approaches enable the identification of the importance and contribution of different features (i.e., contributing factors) to injury severity in traffic accidents, intuitively interpreting the nonlinear behaviors of individual features in traffic accidents with different injury severities.

Therefore, first, we utilize visualization techniques for the

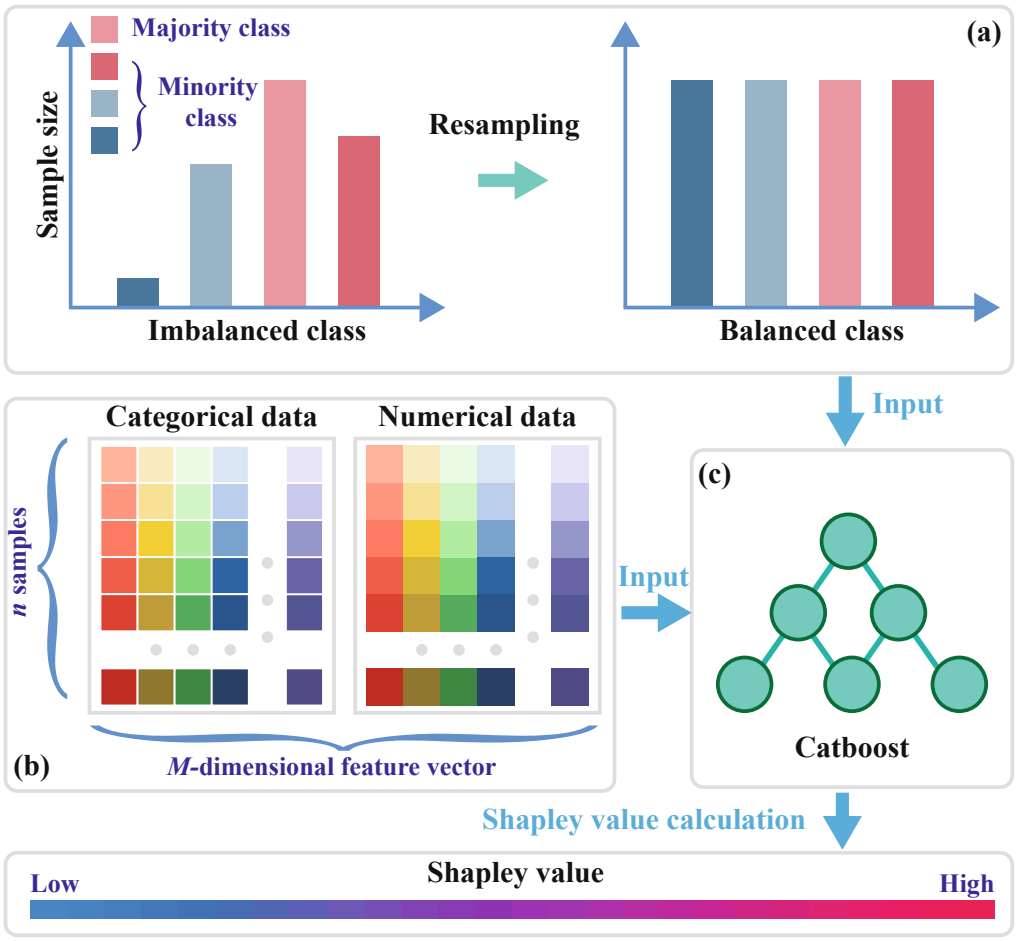

Fig. 2. Features analysis by employing CatBoost combined with the Shapley value.

spatiotemporal feature analysis; second, we perform a simple descriptive statistical analysis for features excluding spatiotemporal features; finally, we utilize machine learning approaches, i.e., CatBoost combined with the Shapley value, for further analysis regarding feature contribution and importance.

In this paper, the CatBoost is first introduced; it was employed to analyze the importance and contribution of contributing factors for injury severities in a traffic accident. It is a machine learning approach based on gradient boosted decision trees (GBDTs), which is similar to other GBDT, and it has the advantage of being utilized to address problems that include heterogeneous features, noisy data, and complex dependencies (Prokhorenkova et al., 2018). Moreover, the CatBoost enables better handling of category features and is thus applicable to datasets that contain a large amount of categorical data (Deng et al., 2020).

We assume a cleaned traffic accident dataset with n samples, $D = \{(x_i, y_i), i = 1, \ldots, n\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ represents M contributing factors for certain indicators, i.e., M-dimensional feature vector with the corresponding output y_i. In this algorithm, the categorical feature will be substituted, as illustrated in Eq. (1) (Samat et al., 2020).

$$x_{\sigma_p k} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j k} = x_{\sigma_p k}\right] y_{\sigma_j} + a \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j k} = x_{\sigma_p k}\right] + a}, \quad k \in (1, M) \tag{1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_s)$ denotes the number of s random permutations of datasets; a denotes the weight of the prior value P.

Furthermore, we interpret the complex nonlinear behavior of different features for the occurrence of accidents by intuitively utilizing the Shapley value and selecting features accordingly. The Shapley value (Shapley, 1953) is calculated by employing game theory and connects optimal credit allocation with local explanations to explain attribute parameters; therefore, it is utilized to explain the outputs of CatBoost. Typically, it is able to describe the importance of M features or evaluate the behavior of features (Hausken, 2020; Štrumbelj and Kononenko, 2014). Recently, the calculation and comparison of Shapley values for various features can be implemented by utilizing the unified framework SHAP (Shapley additional explanation) (Lundberg et al., 2020).

The calculation of Shapley values is described in Eq. (2), where the original feature is replaced with a binary variable $z' \in \{0,1\}^M$ that represents the presence or absence of each feature.

$$g(z') = \phi_0 + \sum_{m=1}^{M} \phi_m z'_m = \text{bias} + \sum \text{contribution of each feature} \tag{2}$$

where $g(z')$ denotes a local surrogate model of the employed CatBoost model. ϕ_m denotes the contribution of the existence of m^{th} feature to the ultimate output.

More specifically, the Shapley value is designed to evaluate the differences of output O obtained after inputting different features. First, the predicted output of the model without the m^{th} feature is calculated, then, the predicted output of the model with the m^{th} feature is calculated, and then the difference is calculated (see Eq. (3)).

$$\phi_m = \sum_{S \subseteq M \setminus \{m\}} \frac{|S|!(M - |S| - 1)!}{M!}\left[O(S \cup \{m\}) - O(S)\right] \tag{3}$$

where S denotes the subset from all M features except for m^{th} feature,

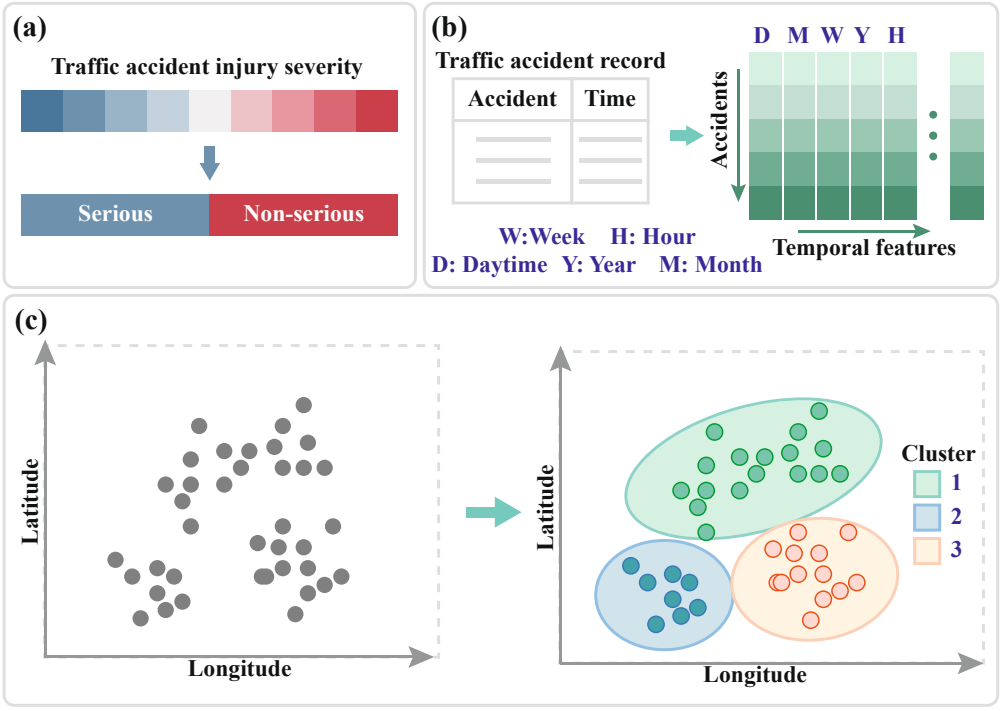

Fig. 3. Data construction. (a) preparation of class labels (b) construction of temporal features (c) clustering according to spatial information.

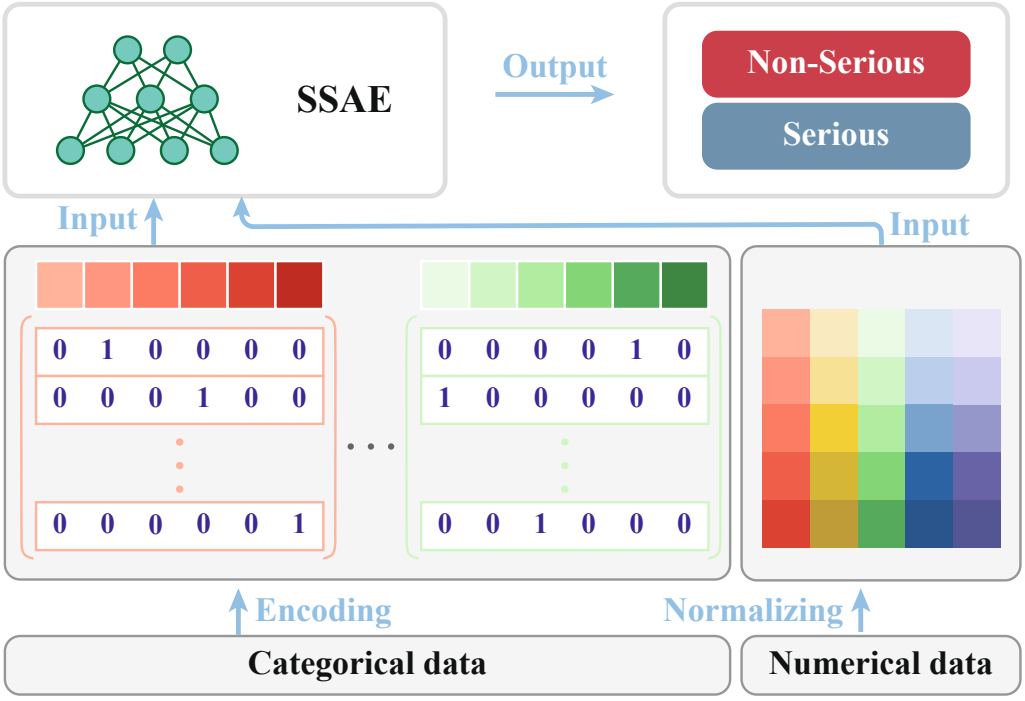

Fig. 4. Application of stacked sparse autoencoder (SSAE) for predicting injury severity in traffic accident.

$O(S)$ denotes the output of CatBoost given the features subset S. $\frac{|S|!(M-|S|-1)!}{M!}$ is the weighting factor to calculate the number of subset permutations. $O(S \cup \{m\}) - O(S)$ denotes the difference that is induced by m^{th} feature.

The feature analysis performed by CatBoost in combination with the shapley value is illustrated in Fig. 2. It is notable that resampling is typically required to resolve imbalance class problems on datasets before input to CatBoost.

2.4. Step 3: Data construction

In this paper, data construction consists of three steps (see Fig. 3). The first step is to set class label for predictive models. Class labels are discrete attributes, which can be predicted based on the values of other contributing factors. For predicting the injury severity in traffic accidents, a common practice is to integrate injury severity into two classes: serious and non-serious. The second is the construction of multiple temporally correlated input features. The third is the utilization of a

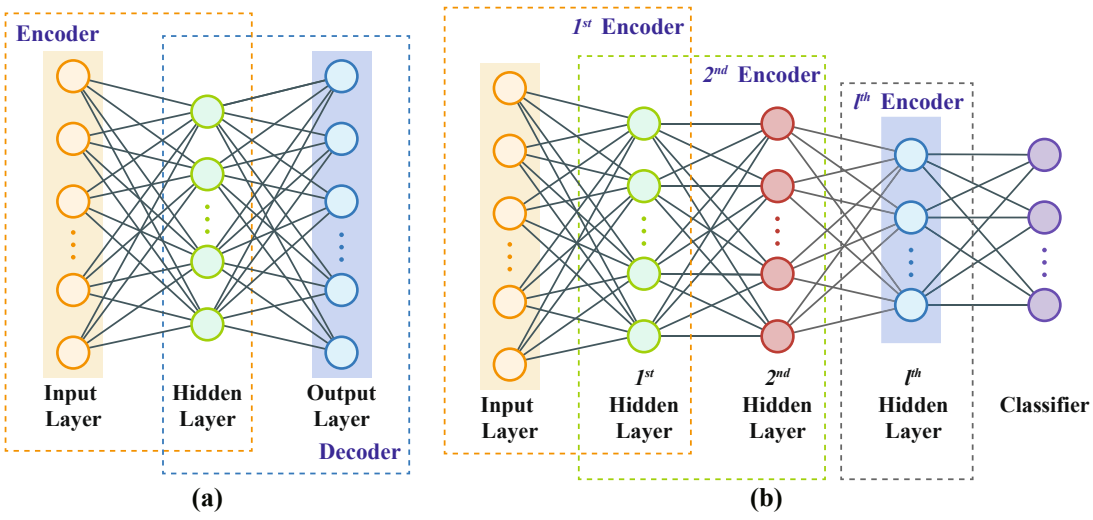

Fig. 5. Model Structure (a) AE (b) Stack AE.

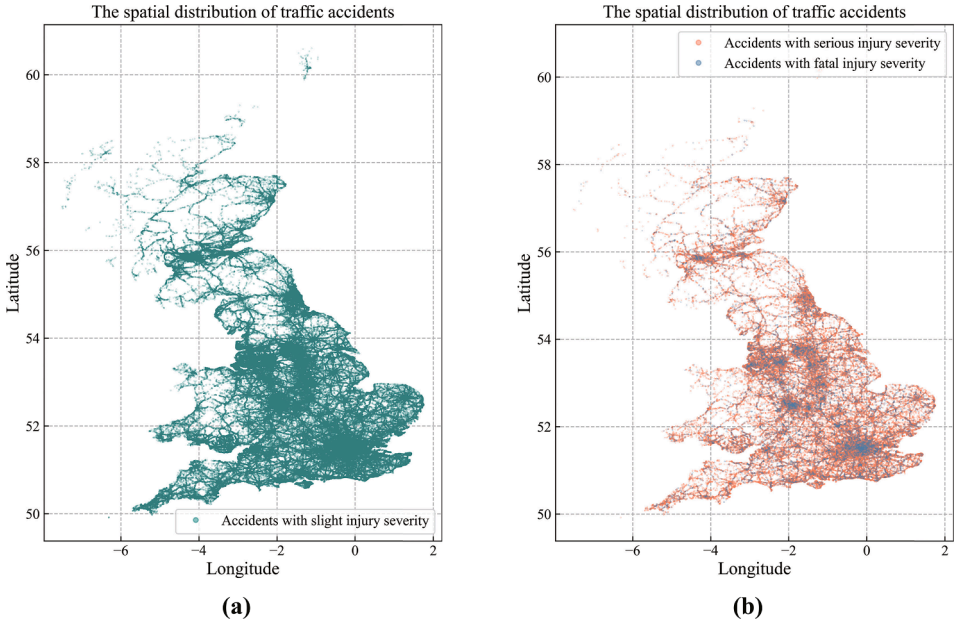

Fig. 6. Spatial distribution of different injury severities.

machine learning approach termed *k*-means clustering to capture the spatial correlation of accident data. According to the location of accidents, the data are divided into several classes to facilitate subsequent modeling separately.

2.5. Step 4: Data modeling

In this section, we introduce details of a novel deep learning model called SSAE, which utilizes the constructed data to predict traffic accident injury severities (i.e., serious or non-serious) (See Fig. 4). First, we briefly review the principles underlying the autoencoder (AE) and the stacked autoencoder architecture; second, we introduce the SSAE, which we employed to extract sparse features.

A single AE (Schmidhuber, 2014) is a typical generative model used in deep learning that consists of three layers, i.e., the input layer, hidden layer, and output layer. These layers constitute an encoder and a decoder, as illustrated in Fig. 5(a). The encoder is designed to transform the input data $x \in \mathbb{R}^d$ to a hidden representation $h \in \mathbb{R}^p$, and the decoder is designed to map the hidden representation to a reconstructed version, $x' \in \mathbb{R}^d$. The process of encoding and decoding is presented in Eqs. (4) and (5).

$$h = f(x) = \sigma\left(W_1^T x + b_1\right) \tag{4}$$

$$x' = g(h) = \sigma\left(W_2^T h + b_2\right) \tag{5}$$

where W_1 denotes the weight matrix of the input layer and the hidden layer; W_2 denotes the weight matrix of the hidden layer and the output layer; b_1 and b_2 are bias vectors; and $\sigma(\cdot)$ refers to the activation function (e.g., Sigmoid, Tanh, and ReLU).

The training objective of AE is to search for the parameters $\theta = W_1, W_2, b_1, b_2$ with the minimized reconstruction loss between the input and output data (see Eq. (6)). To avoid the overfitting, a regularization coefficient λ is commonly required.

$$\arg\min_{\theta} J_{AE}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - x_i'\right\|^2 + \lambda\left(\|W_1\|_2 + \|W_2\|_2\right) \tag{6}$$

Due to the relatively simple and shallow structure, a single AE has a

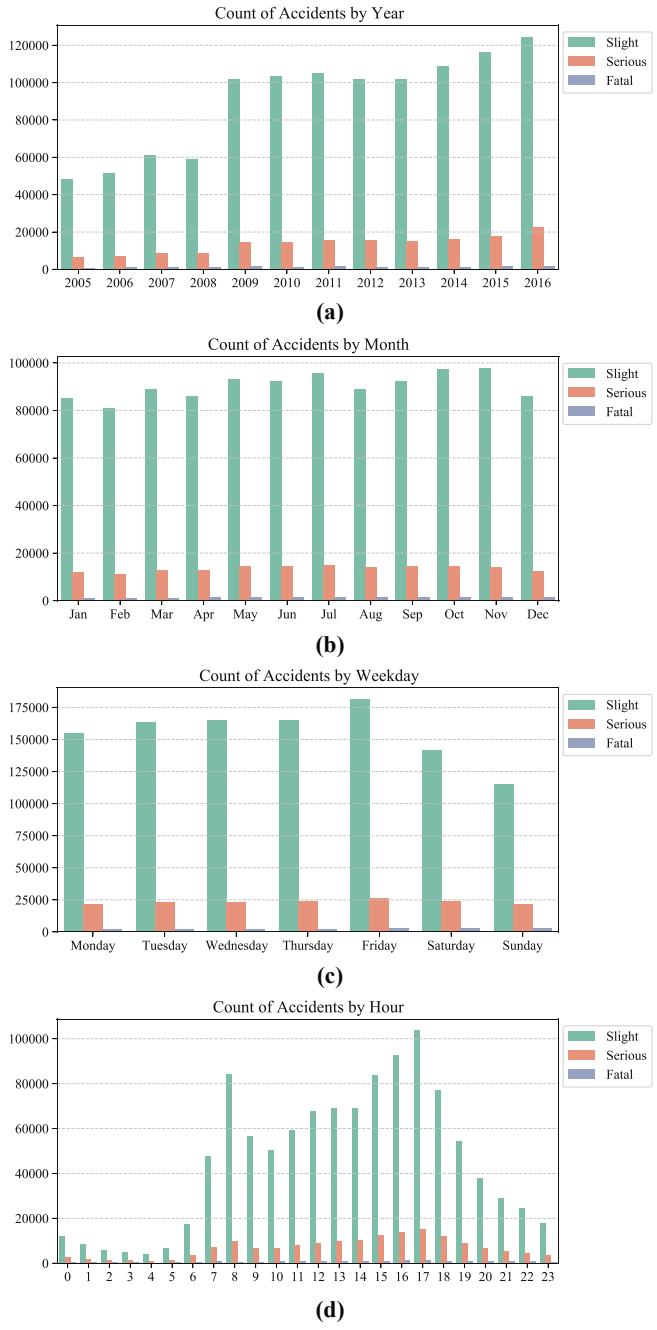

Fig. 7. Temporal distribution of different injury severities by year, month, day of the week, and hour of the day.

limited representative capability. As an improvement, a typical stacked autoencoder architecture is developed by stacking AEs one after another, which utilizes the output of the previous layer of the hidden layer as the input of the previous layer, and the structure is illustrated in Fig. 5b(b) (Liu and Chen, 2019).

Furthermore, although the feature representation learned from the AE is probably perfectly reconstructed using the original inputs, the capability of automatically learning features is lost when there are more hidden layer neurons than input neurons (Pulgar et al., 2020). Therefore, adding a sparsity penalty term to the cost function of an AE provides a more abstract and representative sparse autoencoder than an AE (Ng et al., 2011). A sparse autoencoder is able to compress most of the output of the hidden layer, and an improvement in the training objectives is presented in Eq. (7).

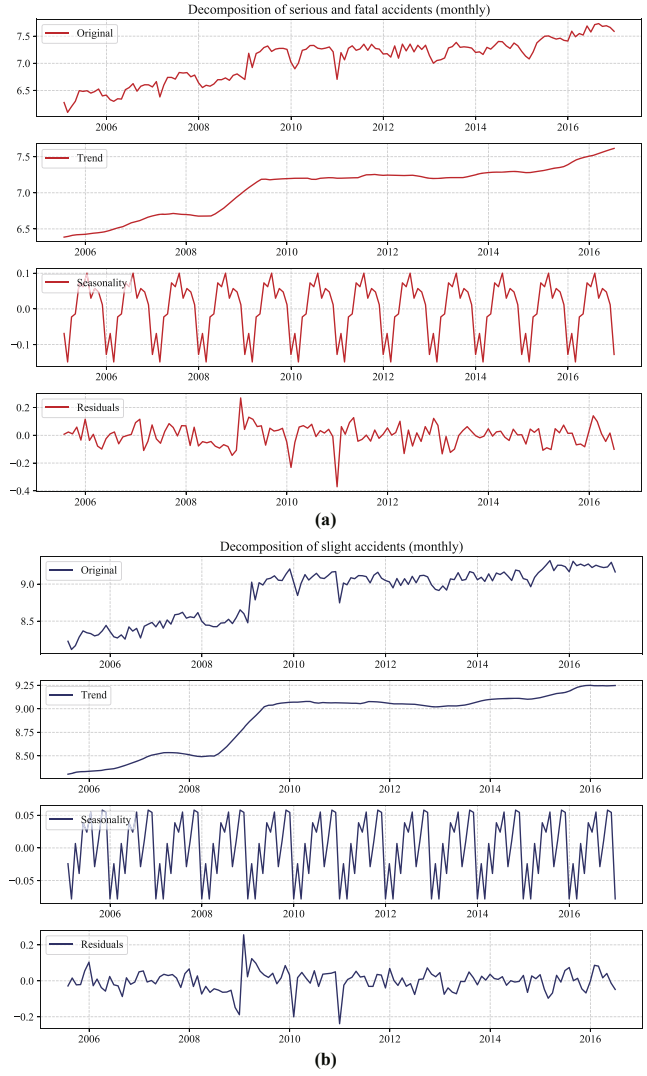

Fig. 8. Temporal pattern of different injury severities by month.

$$\arg\min_{\theta} J_{\text{sparse}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - x_i^{'}\right\|^2 + \lambda\left(\|W_1\|_2 + \|W_2\|_2\right) + \beta\sum_{j=1}^{d} KL\left(\rho\|\widehat{\rho}_j\right) \tag{7}$$

where $\sum_{j=1}^{d} KL\left(\rho\|\widehat{\rho}_j\right)$ denotes the sparsity penalty term, β controls the weight of the sparsity penalty term, $\widehat{\rho}_j$ is the average activation of the jth hidden neuron, ρ is the sparsity parameter, and d denotes the number of hidden neurons.

A SSAE deep neural network is employed by introducing sparse constraints to each AE.

3. Results: A real case

3.1. Data descriptive

To evaluate the effectiveness and applicability of the proposed analytic framework, a UK traffic accident dataset is employed in this work. The dataset between January 2005 and December 2016 is provided by the UK Department for Transport, and can be obtained from an open data source (https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles).

There are two datasets in total, corresponding to accident information and vehicle information, respectively, and each column in the

Table 1
Descriptive statistics for categorical features.

Variables	Category	Serious and Fatal Injury Severity		Slight Injury Severity	
		Freq.	%	Freq.	%
General accident characteristics					
Junction detail	Not at junction or within 20 meters	52,070	46.58	245,081	37.2
	T or staggered junction	33,647	30.1	210,198	31.91
	Crossroads	9665	8.65	70,672	10.73
	Roundabout	5971	5.34	64,323	9.76
	Private drive or entrance	4962	4.44	26,516	4.03
	Other junction	2294	2.05	14,920	2.26
	Slip road	1458	1.3	11,924	1.81
	Mini roundabout	898	0.8	8692	1.32
	More than 4 arms (not roundabout)	824	0.74	6446	0.98
Road type	Single carriageway	87,724	78.47	475,465	72.17
	Dual carriageway	16,974	15.18	115,494	17.53
	Roundabout	4753	4.25	48,490	7.36
	other	2338	2.09	19,323	2.93
Environmental characteristics					
Light conditions	Daylight	79,283	70.92	496,567	75.38
	Darkness	32,506	29.08	162,205	24.62
Urban or rural area	Urban	56,551	50.59	415,198	63.03
	Rural	55,238	49.41	243,574	36.97
Weather conditions	Fine	95,099	85.07	549,861	83.47
	Raining	13,756	12.31	92,639	14.06
	High winds	1595	1.43	8545	1.30
	Fog or mist	762	0.68	4196	0.64
	Snowing	577	0.52	3531	0.54
Road surface conditions	Dry	79,350	70.98	464,083	70.44
	Wet or damp	30,578	27.35	182,579	27.72
	Other	1861	1.66	12,110	1.83
Speed limit	20.0	2130	1.91	12,981	1.97
	30.0	55,447	49.6	400,053	60.73
	40.0	10,997	9.84	64,338	9.77
	50.0	6241	5.58	30,850	4.68
	60.0	27,303	24.42	90,941	13.8
	70.0	9671	8.65	59,609	9.05
Vehicle characteristics					
Vehicle manoeuvre	Going ahead	70,320	62.9	337,289	51.2
	Turning	17,560	15.71	101,163	15.36
	held up	5529	4.95	728,50	11.06
	Slowing	4824	4.32	62,944	9.55
	Moving off	4558	4.08	29,114	4.42
	Overtaking	3641	3.26	20,893	3.17
	Parked	2552	2.28	16,102	2.44
	Changing lane	1429	1.28	10,438	1.58
	Reversing	1376	1.23	7979	1.21
Vehicle type	Car	85,061	76.09	569,678	86.48
	Motorcycle	18,347	16.41	44,823	6.8
	Van	7466	6.68	39,428	5.99
	Bus	671	0.6	4000	0.61
	Other vehicle	196	0.18	728	0.11
	Agricultural vehicle	48	0.04	115	0.02
Driver characteristics					
Driver gender	Male	79,960	71.53	416,383	63.21
	Female	31,829	28.47	242,389	36.79
Driver age	16∼ 20	9880	8.84	55,944	8.49
	21∼ 25	13,790	12.34	85,034	12.91
	26∼ 35	22,603	20.22	150,334	22.82
	36∼ 45	20,497	18.34	132,469	20.11
	46∼ 55	19,521	17.46	113,074	17.16
	56∼ 65	12,562	11.24	65,271	9.91
	66∼ 75	7288	6.52	34,692	5.27
	Over 75	5648	5.05	21,954	3.33

datasets is composed of factors that relate to the circumstances surrounding the accident occur. We selected the temporal and spatial information (i.e., the date and the location) and 14 additional contributing factors for evaluating the proposed analytic framework. After merging the two datasets, we performed the data cleaning for the collected traffic accident record data, for example, removing missing values.

We first analyzed the spatial and temporal features of the data (2005–2016) by employing visualization techniques to investigate the trend and distribution of the data. Then, to further capture more relevant features, we utilize a dataset in a six-year period (2011–2016) to perform descriptive statistics and machine learning features analysis. After being cleaned, the original dataset (2011–2016) comprised 770,561 accident records involving a total of 1,636,286 vehicles and 1,148,267 people and contained multiple variables that are risk factors potentially influencing injury severities. Moreover, accident injury severities are initially categorized into three levels: slight, serious, and fatal. There were 658,772 slight-injury accidents (85.5%), 102,579 serious-injury accidents (13.3%), and 9210 accidents with fatalities (1.2%).

3.2. Feature analysis and selection

3.2.1. Spatiotemporal feature analysis using visualization techniques

The essential contents of spatiotemporal data feature analysis employing descriptive statistics and graphical techniques include (1) the distribution of spatial and temporal variables according to injury severities, (2) the distribution of temporal variables of injury severities, and (3) decomposition of time-series data that are used to examine future trends.

First, we compared the spatial distribution of accidents with different injury severities. As illustrated in Fig. 6, there is obvious significant spatial clustering in the occurrence of accidents, and the density areas of accidents of different severities remain consistent. More accidents occurred in or near major cities, particularly London, Birmingham, the area between Liverpool, Manchester, Leeds and Sheffield, and Newcastle.

Second, we analyze the temporal distribution of accidents with different injury severities. Fig. 7(a) indicates that the number of accidents with different injury severities is increasing annually. Fig. 7(b) indicates that the number of accidents is the highest in October and the lowest in February. Fig. 7(c) reveals that more traffic accidents occur on weekdays than on weekends. Moreover, the highest number of traffic accidents occurred on Friday, far more than on Saturday, in which the fewest traffic accidents occur. Fig. 7(d) illustrates that the peak times for traffic accidents are 8:00 a.m. and 5:00 p.m., corresponding to commuting times.

Third, to further analyze these time-series data and thus improve the knowledge of the trends, we decomposed the time series. Fig. 8(a) illustrates the trend for serious- and fatal-injury accidents, and Fig. 8(b) illustrates the trend for slight-injury accidents. Both indicate significant increasing trends and seasonal trends. For serious- and fatal-injury accidents, the increasing trend is stronger; moreover, accident occurrence appears to peak around the fall and then decrease in the winter. Slight-injury accidents peak twice: in the fall and winter.

Table 2
Descriptive statistics for numerical features.

Variables	Serious and Fatal Injury Severity					Slight Injury Severity				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Age of Vehicle	111,789	7.75	4.86	1	75	658,772	7.59	4.55	1	70
Engine Capacity (CC)	111,789	1634.11	1121.81	48	17,696	658,772	1677.11	927.85	12	27,664
Number of Vehicles	111,789	2.00	1.58	1	67	658,772	2.00	0.83	1	16

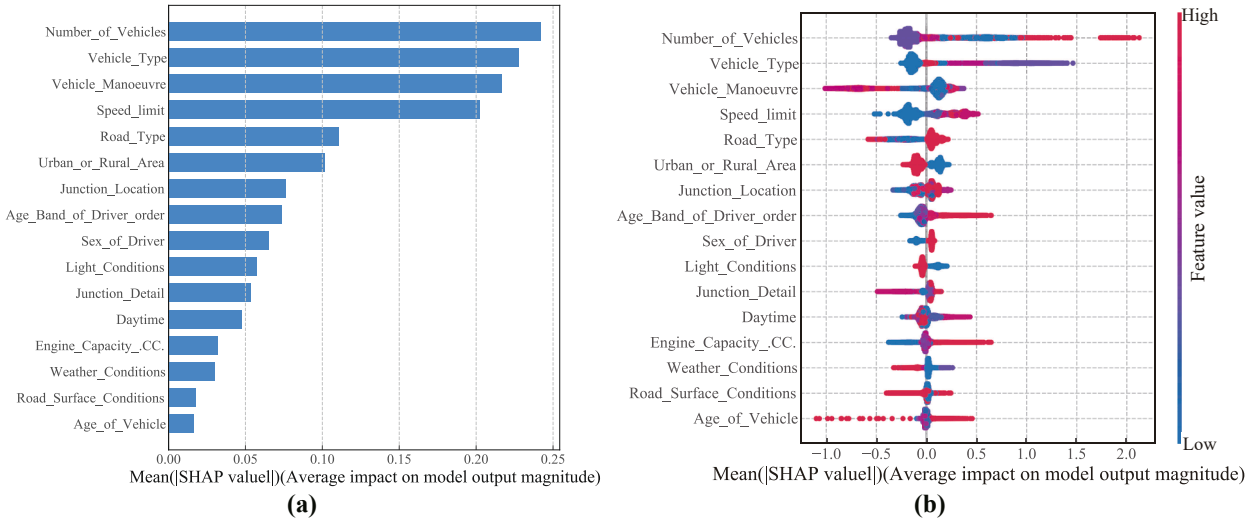

Fig. 9. SHAP summary plot.

3.2.2. *Feature analysis by descriptive statistics*

In this section, we divide the features into numerical and categorical features on the basis of their types and analyze accident injury severity-contributing features by using descriptive statistics.

Table 1 lists the statistical results (i.e., frequencies and percentages) of the categorical features for different injury severities, which consist of 11 features. The features were categorized into four groups: (i) general accident characteristics, including junction details and road types; (ii) environmental characteristics, including the weather, light, road surface conditions, and area type (urban or rural) where the accident occurred; (iii) vehicle characteristics, including the type and maneuver of the vehicle involved in the accident; and (iv) driver characteristics, including the age and gender of the driver.

The statistical description reflects the disadvantages of VRUs in accident safety. For example, the impact of a car on contributing to a serious injury accident is not as significant as the impact of a motorcycle on contributing to a serious injury accident. The car statistically produces similar percentages of serious and non-serious crashes, while motorcycles are responsible for approximately three times as fatal and serious accidents as the slight injury accidents it causes.

Table 2 lists the results of descriptive statistics for the numerical features, including the mean, standard deviation (SD), minimum (min), and maximum (max) values, which consist of 3 features, i.e., the age of the vehicle, engine capacity (CC), and number of vehicles.

3.2.3. *Feature analysis using machine learning approaches*

In this section, we combine Catboost with the Shapley value to analyze and select features. Before training the model, we divide these features into numerical features and categorical features according to the types. To compare the importance of features, for categorical features, we encode it by employing label encoding, i.e., mapping a category with an ordinary number. The data preprocessing consists of three main steps: data splitting, data normalization, and resampling.

First, we employed 75% randomly selected data as the training set and the remaining 25% as the testing set.

Second, for the numerical features in the training and test sets, we standardize each of them by employing the min–max normalization method, respectively. The numerical features in the training set are then combined with the categorical features in the training set to form the training set. Similarly, the numerical features in the test set are combined with the categorical features in the test set to form the test set.

Third, we handle the data imbalance problem that existed in the dataset by employing a synthetic minority oversampling technique (SMOTE), considering that the dataset contains a larger number of slight injury accidents. SMOTE adopts minority data points to generate members that exist between either of the two nearest data points connected on a straight line (Chawla et al., 2002), and has been extensively deployed to handle data imbalance in traffic accident data (Parsa et al., 2019; Parsa et al., 2020).

Finally, when the training process, we employed 10-fold cross-validation on the training data.

Fig. 9 presents a SHAP summary plot that illustrates the importance and contribution of the input features to different injury severities. As shown in Fig. 9(a), according to the results of feature importance ranking, the four features that have the greatest influence on accident injury severity are the vehicles involved in the accident (i.e., number of vehicles), vehicle type, vehicle maneuvering, and speed limit, which implies that some of the vehicle characteristics are important factors affecting accident injury severity. In contrast, environmental characteristics, including weather conditions and road surface conditions, have a relatively less significant influence on accident injury severity. The least influential feature is the age of the vehicle.

Fig. 9(b) further illustrates the range and distribution of the impacts of different features on the injury severity of traffic accidents. The scatter here represents the different Shapley values of the features, and each point is colored by the value of the feature from low (blue) to high (red). The density of the points indicates their distribution in the dataset. For example, the higher number of vehicles is, the higher the Shapley

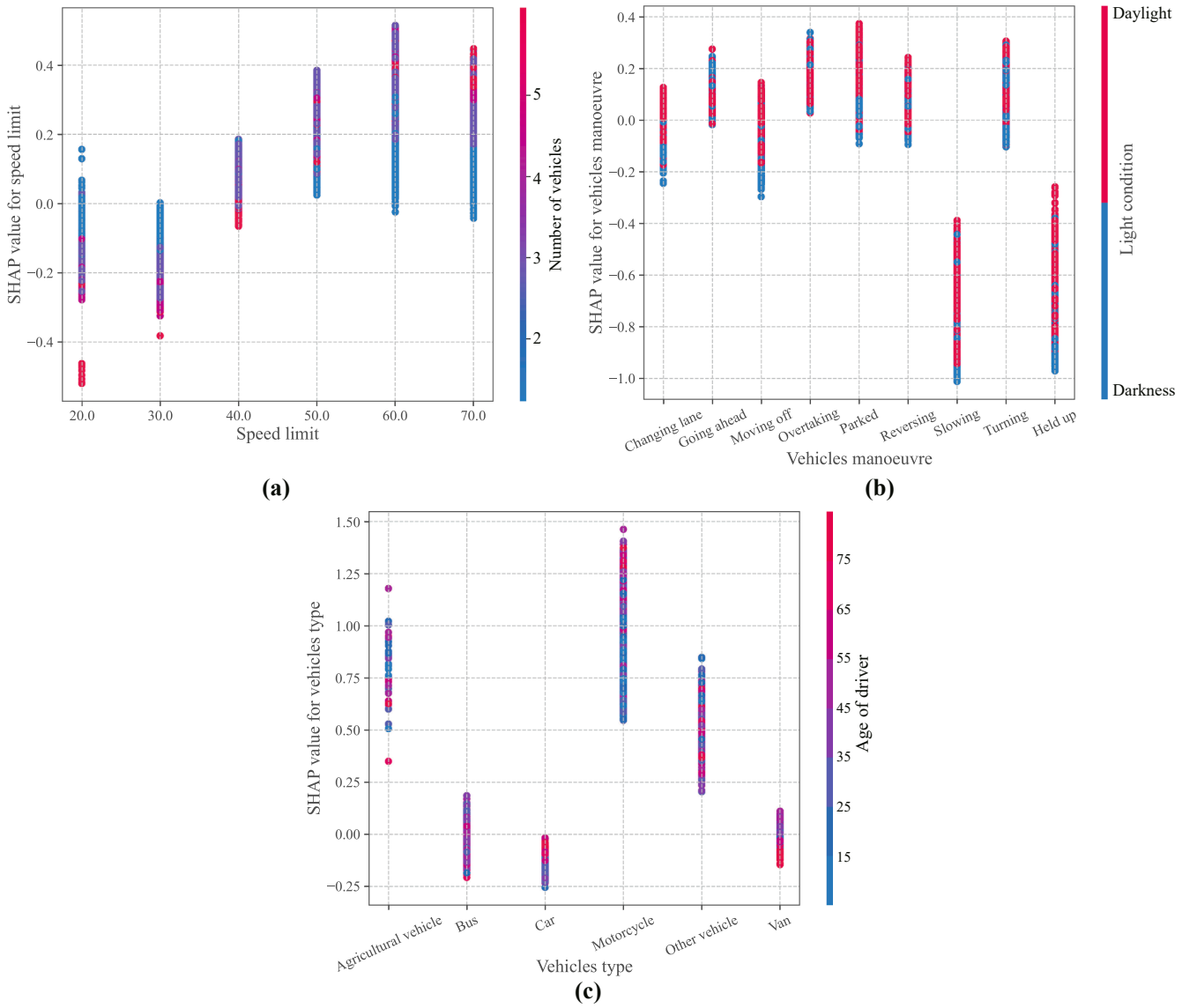

Fig. 10. SHAP dependency analysis.

value and the more significant its influence on contributing to serious accident injuries. A similar trend is observed for the speed limit. These observations are reasonable considering that an increase in the vehicles involved increases the likelihood of fatal injuries; the faster the speed is, the more likely that serious injuries occur (Castillo-Manzano et al., 2019).

The SHAP feature dependency plot is primarily utilized to indicate the distribution and variation of Shapley values with features. For the purpose of interpretation, we analyze the SHAP feature dependency plot by employing three features with relatively great importance.

Fig. 10(a) illustrates the effect of two contributing factors, the speed limit and the number of vehicles. When the speed limit is low, few vehicles are involved in accidents, and more serious accidents are less likely. When the speed limit exceeds 50 mph, the larger the number of vehicles, a more serious injury severity tends to occur. The most prone to serious injury severity is the 60 mph speed limit, where the number of vehicles involved is the average (Wagenaar et al., 1990). This is consistent with the conclusion from previous studies showing that increasing the base speed limit increases the probability of serious-injury accidents (Haleem and Abdel-Aty, 2010; Milton et al., 2008; Khattak and Fontaine, 2020).

Fig. 10(b) illustrates the effect of two contributing factors, the vehicle's maneuver and lighting conditions. Blue represents dark lighting conditions. Obviously, head-on collisions in dark light conditions are prone to lead to more serious injury, whereas slow driving in daylight conditions has a higher association with slight injury. There are various reasons for these results. For example, under dark conditions, the driver's sight distance is presumed to be negatively impacted, and the driver is typically more fatigued when driving at night than when driving during the day (Molan et al., 2020).

Fig. 10(c) illustrates the effect of two contributing factors, the driver's age and the vehicle type. As one of VRUs, the age of motorcyclists contributed more to serious injury severity than the age of car drivers. This means that the age of motorcyclists influences the severity of injuries resulting from accidents more than the age of car drivers. Middle-aged drivers are more prone to be involved in more serious accidents. This is consistent with previous studies and is explainable by the observation that middle-aged drivers are accustomed to driving and commonly exhibit behavior associated with less caution (Adebisi et al., 2019; Zubaidi et al., 2021).

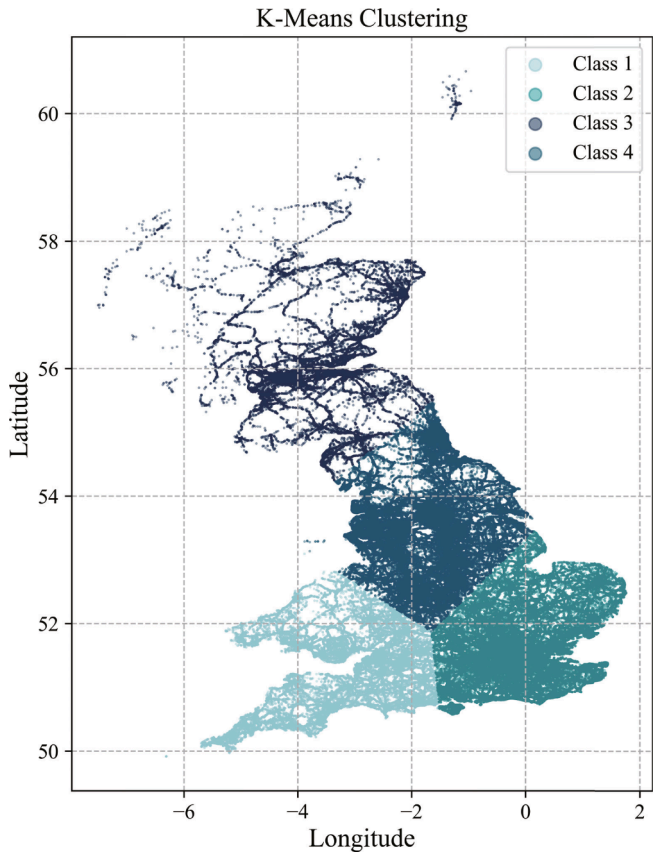

Fig. 11. Clustering results.

Table 3
Confusion matrix.

	Predicted: Non-serious	Predicted: Serious
True: Non-Serious	$T_{non\text{-}serious}$	$F_{serious}$
True: Serious	$F_{non\text{-}serious}$	$T_{serious}$

Table 4
Performance measures of models for a serious injury accident.

Models	Clustering Results	Weighted Average Precision	Weighted Average Recall	Weighted Average F-measure
DNN	Class1	0.71	0.84	0.77
	Class2	0.75	0.87	0.80
	Class3	0.69	0.83	0.75
	Class4	0.72	0.85	0.78
AE	Class1	0.74	0.81	0.77
	Class2	0.76	0.84	0.79
	Class3	0.72	0.81	0.75
	Class4	0.75	0.84	0.78
SSAE	Class1	0.79	0.83	0.80
	Class2	0.80	0.85	0.82
	Class3	0.78	0.82	0.79
	Class4	0.80	0.85	0.84

3.3. Prediction of injury severity

3.3.1. Data construction

Data construction included (1) classification of the three injury severity levels in the dataset into two class labels, i.e., serious/non-serious; (2) construction of temporally correlated data; and (3) clustering based on spatially relevant information. In this section, the dataset adopted (2011–2016) has been cleaned and feature-selected.

First, for the level of traffic accident injury severity, fatal and serious accident injury severities are combined into a class, i.e., the serious level. In contrast, slight crashes are defined as another class, i.e., the non-serious level. Second, we constructed five new temporally correlated features from existing temporal data, including the month, day of the week, hour, time period, and whether the accident occurs on a holiday. Finally, we utilize the *k*-means clustering method to capture the spatial patterns in the large-scale accident record dataset and to categorize the data into four classes according to the spatial clustering results obtained by the elbow method (see Fig. 11).

The *k*-means clustering is a typical machine learning approach. It is centroid-based clustering, which randomly identifies several points from the dataset as the initial cluster centers and assigns each point to the nearest center. Then, it separately calculates the average of each set of parameters to form a new set of updated cluster centers. The process is repeated until no point changes a cluster or the number of iterations achieves a previously specified maximum value. The elbow method enables the plotting of the explained variance as a function of the number of clusters, which allows the selection of the elbow of the curve as the number of clusters (Liu and Deng, 2021).

Furthermore, we perform two preprocessing processes on the employed datasets, including data splitting and data normalization. More specifically, after random splitting, 80% of the data is assigned to training and the remaining 20% to testing. Moreover, for the selected features, we handle the categorical features by utilizing the one-hot coding method and standardize the numerical features with the min–max normalization method.

3.3.2. Evaluation metrics of prediction

We evaluate model performance based on matching matrix values: true positive, false positive, true negative and false negative. A true positive ($T_{Serious}$) occurs when a model correctly identifies a serious injury. A false positive ($F_{Serious}$) occurs when a model wrongly detects a non-serious injury as a serious injury. A false negative ($F_{non\text{-}Serious}$) occurs when a model marks a serious injury as a non-serious injury. A true negative ($T_{non\text{-}Serious}$) occurs when a non-serious injury is correctly classified. These values are constructed based on the confusion matrix in Table 3.

Consequently, we evaluate model performance by employing three metrics, (1) precision, (2) recall, and (3) F-measure (F1), which can be calculated by Eq. (8), Eq. (9), and Eq. (10).

$$\text{Precision} = \frac{T_{serious}}{T_{serious} + F_{serious}} \tag{8}$$

$$\text{Recall} = \frac{T_{serious}}{T_{Serious} + F_{non\text{-}serious}} \tag{9}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.3.3. Predicted results

The experimental setup consisted of (1) an implementation environment, (2) baseline methods, and (3) model hyperparameters. First, we implemented the proposed SSAE model based on the TensorFlow framework (Abadi et al., 2016). Second, we utilized a simple AE and Deep Neural Networks (DNN) as baseline models. Finally, the detailed hyperparameter settings for the model were as follows. The encoder and decoder of the SSAE included a total of four hidden layers with 52, 25, and 16 neurons. These layers were symmetric around the bottleneck. The sparse target was 0.05, and the sparse weight was 0.4. The numbers of hidden layers and neurons of the AE and DNN were similar to those of the SSAE. For the training parameters of the model, we set the learning rate to 0.001, the batch size to 128, and the training epoch to 100, and we employed the Adam optimizer to minimize the loss function during

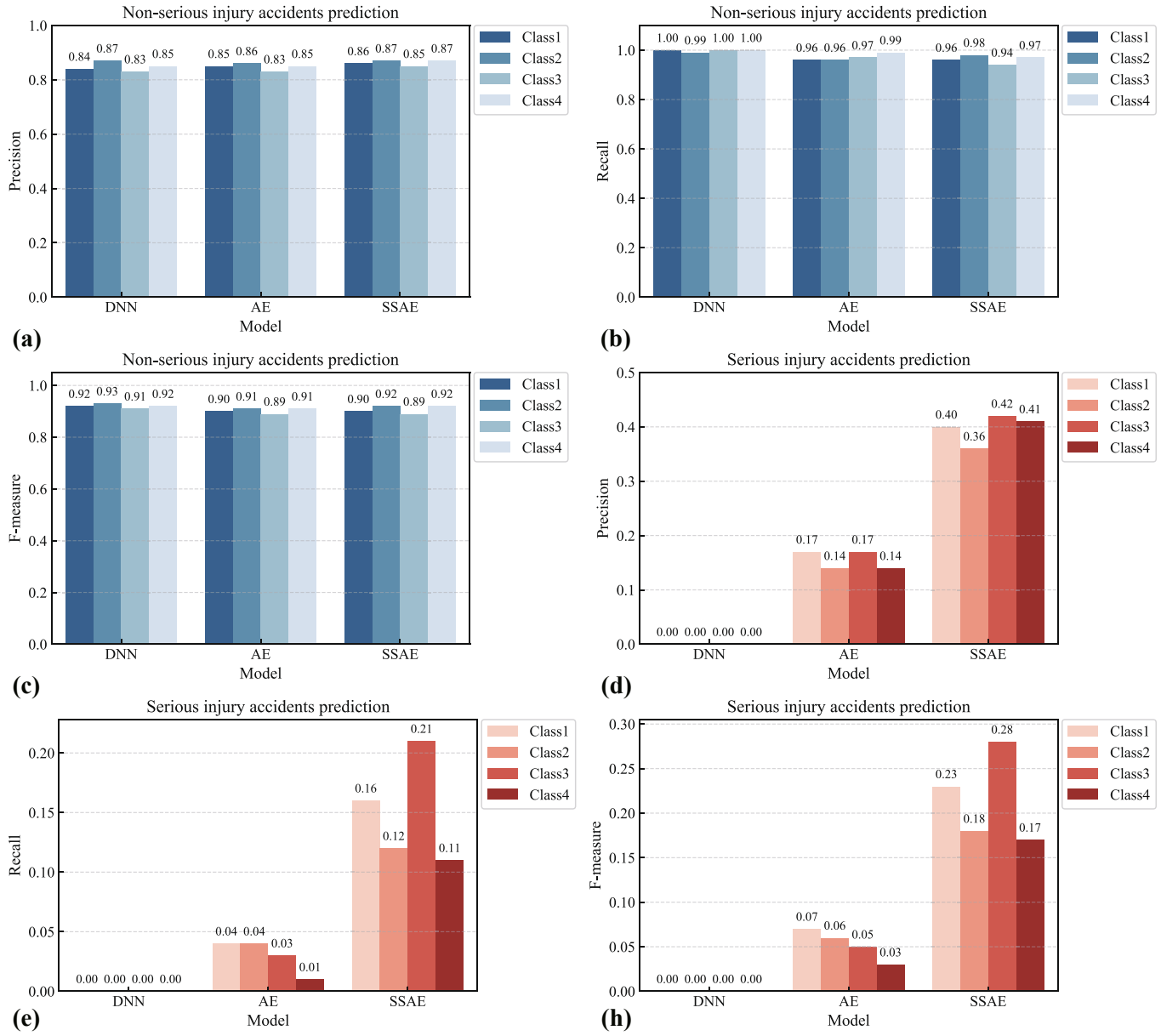

Fig. 12. Non-serious and Serious performance bar plots of precision, recall, and F-measure for the DNN, AE, and SSAE.

the training process. Note that, when models are being trained, 20% of the training dataset was randomly split for validation.

Table 4 illustrates the performance on the test dataset of four different clustered results (i.e., Class1, Class2, Class3, and Class4) under AE, DNN, and SSAE. All models achieve satisfactory performance, which implies these models are sufficient for generalization ability when handling different clustered data. Furthermore, a careful examination of precision and F-measure values is considered necessary when evaluating model performance (Katanalp and Eren, 2020). For all four clustered data, the weighted average precision and weighted average F-measure of SSAE are higher than those of DNN and AE. According to these results, it concludes that the performance of SSAE is better than the other two models. Regarding the performance of SSAE, the best performance is achieved for the datasets Class2 with the weighted average precision and weighted average F-measure of 0.80 and 0.82, respectively. This is followed by the datasets Class4 under SSAE, whose the weighted average precision and weighted average f-measure are 0.80 and 0.84, respectively.

Fig. 12 illustrates the performance for the different models. The results indicate that in predicting non-serious injury, all three models achieve satisfactory performance. In predicting serious accidents, DNN could not classify the serious label, due to its three metric values of 0. SSAE outperforms AE. This means that SSAE is more appropriate for the prediction of serious accidents than AE and DNN, by maximizing the information maintained in the input data.

Furthermore, the performance of the four classes of data differed slightly. More specifically, dataset Class3 outperformed the other datasets in predicting serious injuries with precision, recall, and F-measure of 42%, 21%, and 28%, respectively. In contrast, dataset Class4 performed slightly less well in predicting serious injuries, with recall and F-measure of 11% and 17%, respectively. Considering that the size of dataset Class4 is larger than that of dataset Class3, it implies that the model's prediction results for serious injuries are related to its clustering results than to its data size.

Fig. 13 illustrates the loss of the model for the training and validation data in the four classes of regions. The results reveal that the model training performance is significant, where the models for the dataset Class2 and the dataset Class4 are essentially stable at 40 epochs, while

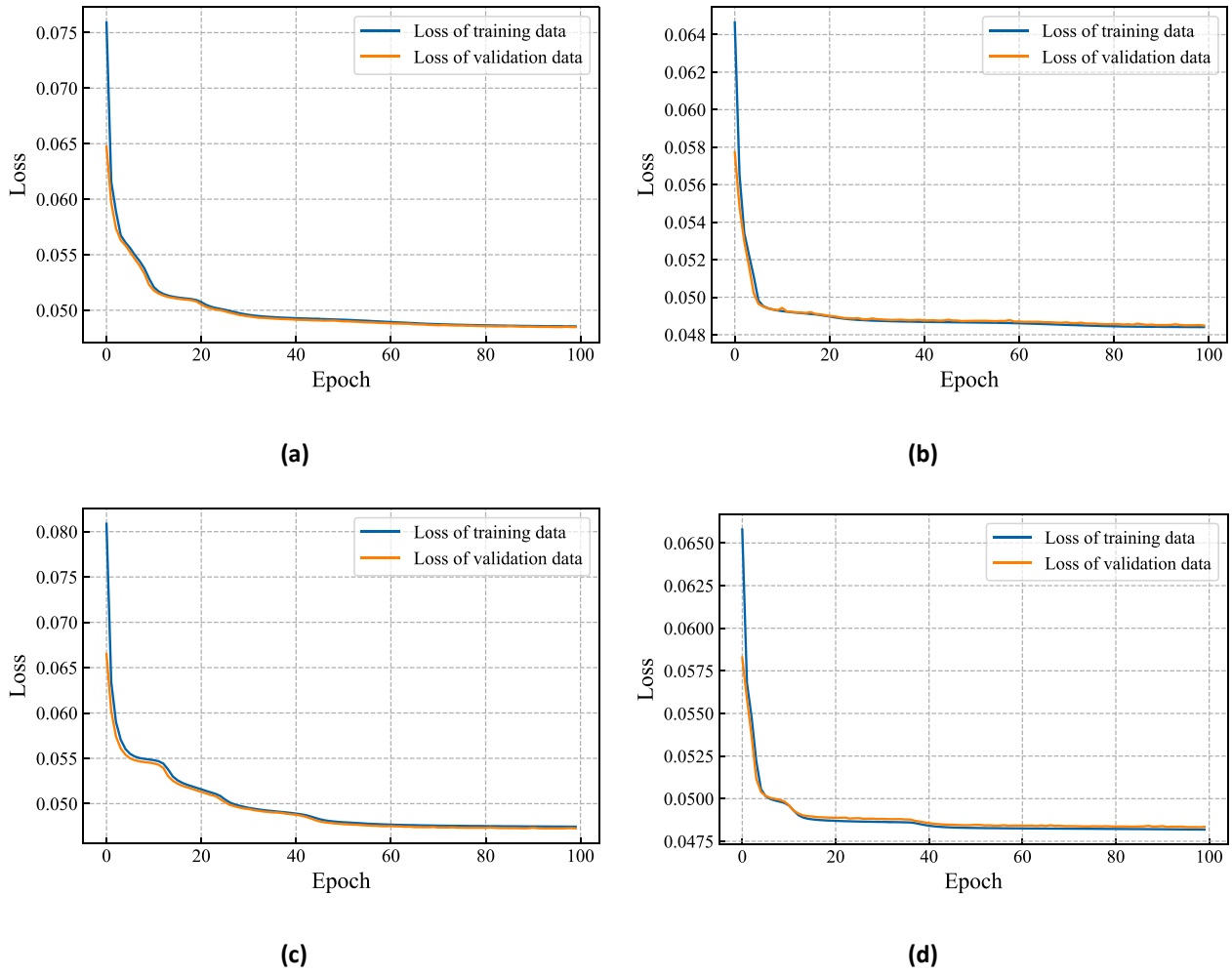

Fig. 13. Loss of training and validation datasets in four clustered data (a) class 1 (b) class 2 (c) class 3 (d) class 4.

the models for the dataset Class1 and the dataset Class3 continue to decrease at 40 epochs and stabilize in 60 epochs.

Fig. 14 illustrates the confusion matrix of the SSAE prediction results for serious and non-serious injuries in all four clustered data. The results indicate that the model predicts accidents with serious injuries. These results are indicators of the effectiveness of the SSAE in predicting injury severity levels of traffic accidents.

4. Discussion

In this paper, an analytic framework using three machine learning approaches, CatBoost, *k*-means clustering, and SSAE, is proposed by considering spatiotemporal correlation and other injury severity contributing factors in traffic accidents. The framework first employs CatBoost to analyze the contributing factors and identify the higher correlated features and then employs *k*-means clustering to cluster the traffic accident data into different classes according to spatial information. Finally, based on the high correlation contributing factors and constructed temporal features, in each clustered data, the framework employs a deep learning model termed SSAE to predict serious/nonserious injuries in traffic accident. The analytical results of the proposed framework demonstrate its effectiveness and applicability.

In this section, we analyze the advantages and shortcomings of the proposed analytical framework. Furthermore, we discuss some potential future work to address these shortcomings.

4.1. Advantages of the proposed method

The advantage and essential idea of the proposed method are the ultilization of different machine learning approaches into the traffic accident injury severity analysis. Each of these machine learning approaches is maximized in the corresponding analysis steps according to their respective strengths. Consequently, an analytic framework for a comprehensive analysis of traffic accident recording data in multiple steps is developed.

First, we analyzed features through the machine learning approach (i.e., CatBoost) combined with the Shapley value. The importance or dependence of various contributing factors revealed as a result is consistent with previous findings from empirical or statistical analysis. This illustrates that this machine learning approach with explanations is available for analyzing contributing factors in accident data as a complement to empirical-based and other statistical methodologies. Furthermore, the method provides a feature importance ranking, which can be adopted to select highly correlative contributing factors for any subsequent modeling.

Second, we employed a machine learning approach (i.e., *k*-means clustering) to classify accident data into different classes depending on their spatial information. This approach considers the spatial pattern of accident occurrence. There is a hypothesis that supports this method of capturing spatial correlation, i.e., accidents that occur on the similarity of certain types of roads and in neighboring areas share certain similarities (Ouni and Belloumi, 2019; Blazquez et al., 2018). Moreover, for subsequent modeling predictions, the advantage of modeling in different

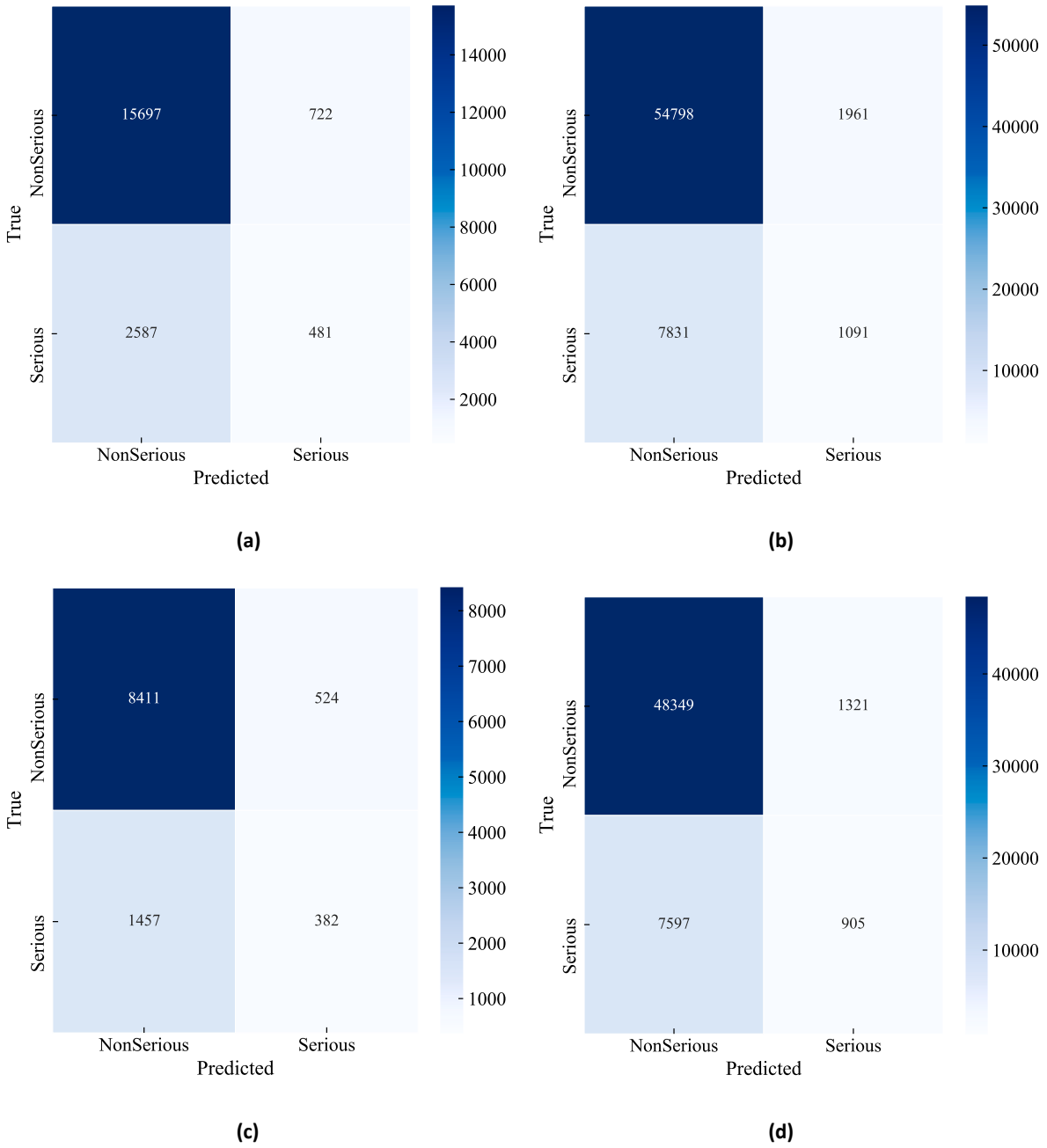

Fig. 14. Normalized confusion matrix of prediction on test data.

spatially-defined classes is a reduction in computational complexity (Li et al., 2021).

Finally, we leveraged the SSAE-based deep learning model, which predicts injury severity in traffic accidents based on high correlation contributing factors. The model reduces feature spaces while further removing redundant and irrelevant contributing factor information in accident data. As an AE-based deep learning model, the model can be considered as a solution to address the imbalanced data problem in traffic accident datasets (Zhao et al., 2021; Islam et al., 2021; Wang et al., 2020).

4.2. Shortcomings of the proposed method

The major limitation of the proposed method is data integrity, considering that the developed framework can be applied for predicting (1) injury severity for traffic accident involving VRU and (2) the injury severity of a VRU in traffic accident. In this paper, the employed dataset lacks information for whether VRUs are involved in a traffic accident, and lacks more detailed characteristics of the VRUs, including the number of bicycles and the number of pedestrians. The absence may render the proposed framework difficult to achieve the desired results. Therefore, datasets containing more VRUs-related information are required, enabling the proposed framework to contribute in alleviating the adverse consequences of traffic accidents for VRU safety.

Moreover, there are several limitations related to the selection of the machine learning approach and the corresponding parameters. For example, common traffic accident records are predominantly categorical data. Generally, CatBoost is appropriate for processing categorical data; therefore, it is employed for analyzing contributing factors in accident data. To improve the accuracy of the analytic framework and make it available for more diverse accident data analysis, a more efficient method for the reasonable selection of optimal machine learning

approaches and corresponding parameters is desirable.

4.3. Outlook and future work

In the future, we plan to apply the proposed framework to datasets that contain more VRUs-related information. Once these appropriate datasets are collected, we can prepare various label classes for prediction targets (i.e., injury severity in traffic accident), for example, serious level involving VRU, non-serious level involving VRU, serious level not involving VRU, and non-serious level not involving VRU.

By leveraging VRU information existing in new datasets, the proposed analytic framework is able to (1) utilize the Catboost combined with the Shapley value to determine which contributing factors have more influence on different class labels (e.g., serious level involving VRU and non-serious level involving VRU) and then make the feature analysis and selection; (2) leverage the proposed deep learning model to perform multi-class prediction by using the chosen features. These analysis and prediction is able to assesse road safety performance, and thus alleviate the adverse consequences of traffic accidents for VRU safety.

On the other hand, we plan to streamline injury severity analysis involving VRUs by leveraging AutoML (He et al., 2021). An AutoML framework is available for implementing self-optimizing modeling for injury severity analysis, which assembles the essential modeling steps into an end-to-end machine learning automation pipeline to obtain high correlation contributing factors, optimal models, and appropriate hyperparameters.

5. Conclusion

In this paper, we propose an analytic framework that employs a deep learning model referred to as SSAE for predicting the injury severity of traffic accidents based on contributing factors. The essential idea of the method is to integrate a single analysis for injury severity of traffic accident into a comprehensive analytic framework and perform the corresponding processing and analysis of accident data by different machine learning approaches. In the proposed method, (1) the contribution of different factors to injury severity is analyzed by utilizing a machine learning approach (i.e., CatBoost) combined with the Shapley value, and low correlation factors are removed, (2) spatially correlated data (i.e., coordinates) of accident records are categorized into different classes by utilizing a machine learning approach (i.e., *k*-means clustering), and (3) the SSAE-based deep learning model is employed to predict serious/nonserious injuries in each data class by using the high correlation factors and the constructed temporal features. The results obtained by analyzing a six-year traffic accident dataset from the UK indicate that (1) the importance and dependence of contributing factors can be obtained by CatBoost and the Shapley value, (2) the SSAE-based deep learning model achieved the best performance compared with the performance of baseline deep learning models. The proposed analytic framework is applicable for other datasets that contain more information related to VRUs, and thus alleviate the adverse consequences of traffic accidents for VRU safety. Future work will focus on improving the performance and efficiency of the analytic framework for model and parameter selection by more automated and flexible methods.

CRediT authorship contribution statement

Zhengjing Ma: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Gang Mei:** Conceptualization, Software, Formal analysis, Investigation, Methodology, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing. **Salvatore Cuomo:** Supervision, Investigation, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgments

This research was jointly supported by the National Natural Science Foundation of China (Grant Nos. 11602235), the Fundamental Research Funds for China Central Universities (2652018091), and the Major Program of Science and Technology of Xinjiang Production and Construction Corps (2020AA002).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. USENIX Association, USA, pp. 265–283.

Abdel-Aty, M., Ekram, A.A., Huang, H., Choi, K., 2011. A study on crashes related to visibility obstruction due to fog and smoke. Accident Analysis and Prevention 43, 1730–1737. https://doi.org/10.1016/j.aap.2011.04.003.

Abellán, J., López, G., De Oña, J., 2013. Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications 40, 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027.

Adebisi, A., Ma, J., Masaki, J., Sobanjo, J., 2019. Age-related differences in motor-vehicle crash severity in california. Safety 5, 1–17. https://doi.org/10.3390/safety5030048.

Ahmad, N., Ahmed, A., Wali, B., Saeed, T.U., 2019. Exploring factors associated with crash severity on motorways in pakistan. In: Proceedings of the Institution of Civil Engineers - Transport 04, pp. 1–20. doi: 10.1680/jtran.18.00032.

Anderson, T., 2009. Kernel density estimation and k-means clustering to profile road accident hotspots. Accident Analysis and Prevention 41, 359–364. https://doi.org/10.1016/j.aap.2008.12.014.

Assi, K., Rahman, S., Mansoor, U., Ratrout, N., 2020. Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. International Journal of Environmental Research and Public Health 17, 1–17. https://doi.org/10.3390/ijerph17155497.

Azeroual, O., Saake, G., Schallehn, E., 2018. Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management 41, 50–56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007.

Bao, J., Liu, P., Ukkusuri, S., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accident Analysis and Prevention 122, 239–254. https://doi.org/10.1016/j.aap.2018.10.015.

Behnood, A., Mannering, F., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: Some empirical evidence. Analytic Methods in Accident Research 8, 7–32. https://doi.org/10.1016/j.amar.2015.08.001.

Bichicchi, A., Belaroussi, R., Simone, A., Vignali, V., Lantieri, C., Li, X., 2020. Analysis of road-user interaction by extraction of driver behavior features using deep learning. IEEE Access 8, 19638–19645. https://doi.org/10.1109/ACCESS.2020.2965940.

Blazquez, C., Picarte, B., Calderón, J., Losada, F., 2018. Spatial autocorrelation analysis of cargo trucks on highway crashes in chile. Accident Analysis and Prevention 120, 195–210. https://doi.org/10.1016/j.aap.2018.08.022.

Castillo-Manzano, J., Castro-Nuño, M., López-Valpuesta, L., Vassallo, F., 2019. The complex relationship between increases to speed limits and traffic fatalities: Evidence from a meta-analysis. Safety Science 111, 287–297. https://doi.org/10.1016/j.ssci.2018.08.030.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research (JAIR) 16, 321–357. https://doi.org/10.1613/jair.953.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of Safety Research 40, 317–327. https://doi.org/10.1016/j.jsr.2009.05.003.

De Oña, J., López, G., Abellán, J., 2013a. Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention 50, 1151–1160. https://doi.org/10.1016/j.aap.2012.09.006.

De Oña, J., López, G., Mujalli, R., Calvo, F., 2013b. Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks. Accident Analysis and Prevention 51, 1–10. https://doi.org/10.1016/j.aap.2012.10.016.

Deng, K., Zhang, X., Cheng, Y., Zheng, Z., Jiang, F., Liu, W., Peng, J., 2020. A remaining useful life prediction method with long-short term feature processing for aircraft engines. Applied Soft Computing Journal 93, 1–10. https://doi.org/10.1016/j.asoc.2020.106344.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accident Analysis and Prevention 82, 192–198. https://doi.org/10.1016/j.aap.2015.05.018.

Dong, S., Zhou, J., 2020.A comparative study on drivers' stop/go behavior at signalized intersections based on decision tree classification model. Journal of Advanced Transportation 2020, 1–13. doi: 10.1155/2020/1250827.
Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40, 1033–1054. https://doi.org/10.1016/j.aap.2007.11.010.
Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., Yuan, J., 2020. Predicting real-time traffic conflicts using deep learning. Accident Analysis and Prevention 136, 1–14. https://doi.org/10.1016/j.aap.2019.105429.
Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of Safety Research 41, 347–357. https://doi.org/10.1016/j.jsr.2010.04.006.
Hausken, K., 2020. The shapley value of coalitions to other coalitions. Humanities and Social Sciences Communications 7, 104–114. https://doi.org/10.1057/s41599-020-00586-9.
He, X., Zhao, K., Chu, X., 2021. Automl: A survey of the state-of-the-art. Knowledge-Based Systems 212, 1–27. https://doi.org/10.1016/j.knosys.2020.106622.
Huang, T., Wang, S., Sharma, A., 2020. Highway crash detection and risk estimation using deep learning. Accident Analysis and Prevention 135, 1–11. https://doi.org/10.1016/j.aap.2019.105392.
Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis and Prevention 108, 27–36. https://doi.org/10.1016/j.aap.2017.08.008.
Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. Accident Analysis and Prevention 151, 1–13. https://doi.org/10.1016/j.aap.2020.105950.
Jiang, F., Yuen, K., Lee, E., 2020. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accident Analysis and Prevention 141, 1–14. https://doi.org/10.1016/j.aap.2020.105520.
Kamruzzaman, M., Debnath, A., Bourdaniotis, V., 2019. An exploratory study on the safety effects of speed limit reduction policy in brisbane and melbourne cbds. In: Australasian Transport Research Forum, ATRF 2019 - Proceedings, pp. 1–12.
Katanalp, B., Eren, E., 2020. The novel approaches to classify cyclist accident injury-severity: Hybrid fuzzy decision mechanisms. Accident Analysis and Prevention 144, 1–17. https://doi.org/10.1016/j.aap.2020.105590.
Khattak, Z., Fontaine, M., 2020. A bayesian modeling framework for crash severity effects of active traffic management systems. Accident Analysis and Prevention 145, 1–10. https://doi.org/10.1016/j.aap.2020.105544.
Kim, J.K., Kim, S., Ulfarsson, G., Porrello, L., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accident Analysis and Prevention 39, 238–251. https://doi.org/10.1016/j.aap.2006.07.002.
Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on lstm-cnn. Accident Analysis and Prevention 135, 1–9. https://doi.org/10.1016/j.aap.2019.105371.
Li, Y., Karim, M., Qin, R., Sun, Z., Wang, Z., Yin, Z., 2021. Crash report data analysis for creating scenario-wise, spatio-temporal attention guidance to support computer vision-based perception of fatal crash risks. Accident Analysis and Prevention 151, 1–13. https://doi.org/10.1016/j.aap.2020.105962.
Liu, F., Deng, Y., 2021. Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems 29, 986–995. https://doi.org/10.1109/TFUZZ.2020.2966182.
Liu, H., Chen, C., 2019. Multi-objective data-ensemble wind speed forecasting model with stacked sparse autoencoder and adaptive decomposition-based error correction. Applied Energy 254, 1–18. https://doi.org/10.1016/j.apenergy.2019.113686.
Liu, Z., Chen, H., Li, Y., Zhang, Q., 2020. Taxi demand prediction based on a combination forecasting model in hotspots. Journal of Advanced Transportation 2020, 1–13. https://doi.org/10.1155/2020/1302586.
Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
Ma, X., Xing, Y., Lu, J., 2018. Causation analysis of hazardous material road transportation accidents by bayesian network using genie. Journal of Advanced Transportation 2018, 1–12. https://doi.org/10.1155/2018/6248105.
Macpherson, A., To, T., Parkin, P., Moldofsky, B., Wright, J., Chipman, M., Macarthur, C., 2004. Urban/rural variation in children's bicycle-related injuries. Accident Analysis and Prevention 36, 649–654. https://doi.org/10.1016/S0001-4575(03)00086-1.
Mauro, R., De Luca, M., Dell'Acqua, G., 2013. Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. Modern Applied Science 7, 11–19. https://doi.org/10.5539/mas.v7n10p11.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis and Prevention 40, 260–266. https://doi.org/10.1016/j.aap.2007.06.006.
Molan, A., Moomen, M., Ksaibati, K., 2020. Estimating the effect of geometric features of side traffic barriers on crash severity of interstate roads in wyoming. Accident Analysis and Prevention 144, 1–10. https://doi.org/10.1016/j.aap.2020.105639.
Mujalli, R., De Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using bayesian networks. Journal of Safety Research 42, 317–326. https://doi.org/10.1016/j.jsr.2011.06.010.
Ng, A., et al., 2011. Sparse autoencoder. CS294A Lecture notes 72, 1–19.
Osama, A., Sayed, T., 2017. Macro-spatial approach for evaluating the impact of socio-economics, land use, built environment, and road facility on pedestrian safety. Canadian Journal of Civil Engineering 44, 1036–1044. https://doi.org/10.1139/cjce-2017-0145.
Ouni, F., Belloumi, M., 2019. Pattern of road traffic crash hot zones versus probable hot zones in tunisia: A geospatial analysis. Accident Analysis and Prevention 128, 185–196. https://doi.org/10.1016/j.aap.2019.04.008.
Parsa, A., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A., 2020. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. Accident Analysis and Prevention 136, 1–8. https://doi.org/10.1016/j.aap.2019.105405.
Parsa, A., Taghipour, H., Derrible, S., Mohammadian, A., 2019. Real-time accident detection: Coping with imbalanced data. Accident Analysis and Prevention 129, 202–210. https://doi.org/10.1016/j.aap.2019.05.014.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: Unbiased boosting with categorical features, in: Curran Associates Inc., Curran Associates Inc. pp. 6639–6649. doi: 10.1016/j.aap.2011.04.003.
Pulgar, F., Charte, F., Rivera, A., del Jesus, M., 2020. Choosing the proper autoencoder for feature fusion based on data complexity and classifiers: Analysis, tips and guidelines. Information Fusion 54, 44–60. https://doi.org/10.1016/j.inffus.2019.07.004.
Riveiro, M., Lebram, M., Elmer, M., 2017. Anomaly detection for road traffic: A visual analytics framework. IEEE Transactions on Intelligent Transportation Systems 18, 2260–2270. https://doi.org/10.1109/TITS.2017.2675710.
Samat, A., Li, E., Du, P., Liu, S., Miao, Z., Zhang, W., 2020. Catboost for rs image classification with pseudo label support from neighbor patches-based clustering. IEEE Geoscience and Remote Sensing Letters 99, 1. https://doi.org/10.1109/LGRS.2020.3038771.
Schmidhuber, J., 2014. Deep learning in neural networks: An overview. Neural Networks 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
Shapley, L., 1953. A value for n-persons games. Annals of Mathematics Studies 28, 307–318.
Shi, X., Wong, Y., Li, M.F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on xgboost for driving assessment and risk prediction. Accident Analysis and Prevention 129, 170–179. https://doi.org/10.1016/j.aap.2019.05.005.
Smith, A., 2016. Spring forward at your own risk: Daylight saving time and fatal vehicle crashes. American Economic Journal: Applied Economics 8, 65–91. https://doi.org/10.1257/app.20140100.
Taamneh, M., Alkheder, S., Taamneh, S., 2017. Data-mining techniques for traffic accident modeling and prediction in the united arab emirates. Journal of Transportation Safety and Security 9, 146–166. https://doi.org/10.1080/19439962.2016.1152338.
Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x.
Vanlaar, W., Mainegra Hing, M., Brown, S., McAteer, H., Crain, J., McFaull, S., 2016. Fatal and serious injuries related to vulnerable road users in canada. Journal of Safety Research 58, 67–77. https://doi.org/10.1016/j.jsr.2016.07.001.
Wagenaar, A., Streff, F., Schultz, R., 1990. Effects of the 65 mph speed limit on injury morbidity and mortality. Accident Analysis and Prevention 22, 571–585. https://doi.org/10.1016/0001-4575(90)90029-K.
Wali, B., Khattak, A., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Analytic Methods in Accident Research 28, 1–38. https://doi.org/10.1016/j.amar.2020.100136.
Wang, Y.R., Sun, G.D., Jin, Q., 2020. Imbalanced sample fault diagnosis of rotating machinery using conditional variational auto-encoder generative adversarial network. Applied Soft Computing Journal 92, 1–19. https://doi.org/10.1016/j.asoc.2020.106333.
Xu, C., Ding, Z., Wang, C., Li, Z., 2019. Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. Journal of Safety Research 71, 41–47. https://doi.org/10.1016/j.jsr.2019.09.001.
Yan, X., Ma, M., Huang, H., Abdel-Aty, M., Wu, C., 2011. Motor vehicle-bicycle crashes in beijing: Irregular maneuvers, crash patterns, and injury severity. Accident Analysis and Prevention 43, 1751–1758. https://doi.org/10.1016/j.aap.2011.04.006.
Yasmin, S., Eluru, N., Bhat, C., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Analytic Methods in Accident Research 1, 23–38. https://doi.org/10.1016/j.amar.2013.10.002.
Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science 63, 50–56. https://doi.org/10.1016/j.ssci.2013.10.012.
Zeng, Q., Huang, H., 2014. A stable and optimized neural network model for crash injury severity prediction. Accident Analysis and Prevention 73, 351–358. https://doi.org/10.1016/j.aap.2014.09.006.
Zhao, Y., Hao, K., Tang, X.S., Chen, L., Wei, B., 2021. A conditional variational autoencoder based self-transferred algorithm for imbalanced classification. Knowledge-Based Systems 218, 1–10. https://doi.org/10.1016/j.knosys.2021.106756.
Zubaidi, H., Obaid, I., Alnedawi, A., Das, S., 2021. Motor vehicle driver injury severity analysis utilizing a random parameter binary probit model considering different types of driving licenses in 4-legs roundabouts in south australia. Safety Science 134, 1–10. https://doi.org/10.1016/j.ssci.2020.105083.