A novel feature extraction model for traffic injury severity and its application to Fatality Analysis Reporting System data analysis

Lixin Yan¹, Yi He²,³,⁴, Lingqiao Qin⁵, Chaozhong Wu²,³, Dunyao Zhu²,⁵ and Bin Ran⁶

¹School of Transportation and Logistics, East China Jiaotong University, Nanchang, P.R. China
²Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan, P.R. China
³Engineering Research Center for Transportation Safety, Ministry of Education, Wuhan, P.R. China
⁴National Engineering Laboratory for Transportation Safety & Emergency Informatics, Beijing, P.R. China
⁵Wuhan KOTEI Informatics Co. Ltd., Wuhan, P.R. China
⁶Transportation Engineering Laboratory, University of Wisconsin-Madison, Madison, WI, USA

Abstract
The prevention of severe injuries during crashes has become one of the leading issues in traffic management and transportation safety. Identifying the impact factors that affect traffic injury severity is critical for reducing the occurrence of severe injuries. In this study, the Fatality Analysis Reporting System data are selected as the dataset for the analysis. An algorithm named improved Markov Blanket was proposed to extract the significant and common factors that affect crash injury severity from 29 variables related to driver characteristics, vehicle characteristics, accidents types, road condition, and environment characteristics. The Pearson correlation coefficient test is applied to verify the significant correlation between the selected factors and traffic injury severity. Two widely used classification algorithms (Bayesian networks and C4.5 decision tree) were employed to evaluate the performance of the proposed feature selection algorithm. The calculation result of the correlation coefficient, accuracy of classification, and classification error rate indicated that the improved Markov Blanket not only could extract the significant impact factors

Corresponding author:
Yi He, Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan 430063, P.R. China.
Email: heyi1201@foxmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
but could also improve the accuracy of classification. Meanwhile, the relationship between five selected factors (atmospheric condition, time of crash, alcohol test result, crash type, and driver’s distraction) and traffic injury severity was also analyzed in this study. The results indicated that crashes occurred in bad weather condition (e.g. fog or worse), in night time, in drunk driving, in crash type of single driver, and in distracted driving, which are associated with more severe injuries.

**Keywords**
Feature extraction, improved Markov Blanket, traffic injury severity, Bayesian networks, C4.5 decision tree

**Introduction**

Road traffic accidents have become the key cause of death of population in the current years. The occurrence of crash, especially fatality crashes, not only caused an immense damage to the humans and the economy but also hindered the development of the society. The statistical analysis results from the Fatality Analysis Reporting System (FARS) in the United States (see Figure 1) presented that the number of crashes was still at high levels (average 50,000 records were collected from 2010 to 2014 in FARS). To be more precise, due to the specificity of the FARS data, a crash was recorded if a person died within 30 days of road traffic accidents, for last 5 years. It has been observed that the records of fatality injury possessed nearly 50% of the total number of records in FARS, and the highest number appeared in 2012, which reached 29,867. Although it seems that the numbers of fatality injury records have decreased in recent years (approximately 17% reduction from 2011 to 2014), considering the damage of the fatality crashes, more efforts are needed to prevent the occurrence of crashes, especially fatality crashes.

Previous study has demonstrated a fatality crash is caused by five types of factors: driver characteristics, vehicle characteristics, accidents types, road condition, and environment characteristics. Many factors such as the length of entire ramps, DeLength (length of deceleration lanes), and others had already been considered

![Figure 1. The statistical analysis for crashes.](image-url)
as feature variables to analyze the crashes or fatality crashes. However, there are 180 variables related to the occurrence of crashes in FARS. Therefore, it is necessary to summarize the features which can provide a comprehensive evaluation of the occurrence of fatality crashes. Feature selection is a useful approach to reduce the dimensionality of feature space. With the development of data collection and machine learning techniques, feature selection plays a critical role in data mining and machine learning. It thus often provides a comprehensive overview to conduct the FARS data.

The scope of this study is to propose an algorithm for extracting impact factors that are significantly related to the occurrence of crash according to their injury severity. Moreover, the quantitative relationship between the injury severity and the selected features was also in-depth analyzed in this study. The purpose of this article is to explore the main causes that result in the severe injury in the United States. These achievements can be associated to improve the time of emergency response and drivers’ training pattern.

Literature review

This review focuses on the research developments of risk factors associated with traffic injury severity, feature selection methods in the field of traffic safety, and classification algorithm for the prediction of traffic injury severity. The constructions are stated in the following sections.

Risk factors associated with traffic injury severity

The risk factors for human, vehicle, road condition, and environment are associated with crashes and injuries/fatalities. Human factors play a critical role in road traffic accidents. As an example, driver fatigue is regarded as one of the most serious reasons which leads to fatal accident and injuries, causing 15%–20% of all traffic accidents in developed countries. Meanwhile, Ma et al. found that alcohol usage and drivers having a license or not significantly affect injury severity. Moreover, the relationship between the injury severity and other human characteristics, such as driver’s sex, age, and emotion, also had been discussed in previous studies.

Meanwhile, Michalaki et al. found that the number of vehicles involved in an accident, low visibility, and the hour of traffic accident had positively affected the injury severity. Abu-Zidan and Eid focused on exploring the risk factors that affect injury severity of vehicle occupants. According to their research, the mechanism of injury, age of the vehicle occupant, and vehicle speed significantly affect the occurrence of injuries/fatalities. In addition, the geometric design variables (curvature, lane width, vertical grade, and so on), traffic conditions (speed limit, average speed, traffic flow, and congestion), weather condition (wind, rain, and fog), and vehicle (vehicle type, wind dynamic, and shape) also had already been confirmed as impact factors that cause traffic injuries/fatalities.
**Feature selection and classification algorithm for the prediction of traffic injury severity**

Feature selection not only can extract the impact factors that significantly affect traffic injury severity but also can improve the accuracy of traffic accident prediction. Two typical feature selection methods, feature ranking and feature subset selection, are widely used to conduct the big data. The feature ranking methods take features as the evaluation units and rank them according to their discrimination power. The feature subset selection methods try to find the best feature subset according to the best discrimination power. As an example, Yang investigated the feature problem for traffic congestion prediction. According to their research, a feature ranking selection method was employed to extract traffic congestion–related features from the multi-sensors signals. In another example, Zhang et al. proposed a hybrid feature selection algorithm named SRSF to overcome the impacts of traffic flows. In addition, other feature selection methods such as rough set, genetic algorithm, correlation-based and causal feature selection, and fast correlation based filter (FCBF) are also employed in the area of traffic safety. However, few feature subset selection methods are used to overcome the impacts of traffic injury severity.

The traffic injury severity classification algorithms can predict the occurrence of traffic accident with high accuracy; moreover, it can also be used to analyze the relationship between impacts and injury severity. Bayesian Networks (BN) is a commonly used algorithm to predict traffic injury severity. Huang et al. proposed a hierarchical Bayesian binomial logistic model to identify the severity level of injury at signalized intersections. Yu and Abdel-Aty developed a hierarchical Bayesian BP model to analyze the level of crash injury severity with real-time traffic data. Meanwhile, the application of classification and regression tree (CART) has been reported to be a powerful tool to analyze traffic safety problems. Chang and Wang established a traffic severity classification model based on CART, and the results indicate that the vehicle type, motorcycle and bicycle riders, and pedestrians are the most important variables for injury severity. Kashani and Mohaymany applied CART model to predict the crash severity on two-lane, two-way rural roads in Iran. In addition, other algorithms, such as artificial network, fault tree analysis (FTA), logistic regression, and Naïve Bayes (NB) have also been adopted to identify traffic crashes and traffic injury severity.

**Methodologies**

**Feature selection algorithm (Markov Blanket)**

Markov Blanket is a typical feature subset selection algorithm, and this is the first time that the algorithm has been used to select attributes in the domain which was put forward by Koller and Sahami. The Markov Blanket has the advantage of being able to extract attributes from data with multi-dimensional high sample size and select attributes by eliminating redundant variables. It has been proved suitable
for traffic data mining in previous studies. Thus, we try to introduce the Markov Blanket theory into the area of feature selection with respect to traffic injury severity, and the definition of this algorithm is described as follows.

The Markov Blanket of the dependent variable $T$, denoted as $MB(T)$, is a minimal set of variables (or factors, features; hereafter, we use these terms interchangeably) conditioned on which all other variables are probabilistically independent of $T$ (Definition 1). Thus, knowing the values of $MB(T)$ is sufficient to determine the probability distribution of $T$, and the values of all other attributes become superfluous. Obviously, we can only use attributes in $MB(T)$ instead of all the attributes for optimal prediction. Moreover, under certain conditions (faithfulness to a BN), $MB(T)$ is the subset that contains parents, children, and parents of children of the target $T$ in the BN.

**Definition 1 (Markov Blanket).** The Markov Blanket of a target attribute $T \in V$, denoted as $MB(T)$, is a minimal subset of attributes for which

$$(T \perp V - MB(T) - T|MB(T))$$

where $V$ is the set of all attributes in the domain, and symbol “$\perp$” denotes independence. The specific process of the MB algorithm is as follows.

---

**Algorithm 1. MB algorithm**

**Input:** Training dataset $D (F, T)$ /* $F$ is the factor set and $T$ is the target */

**Output:** $MB(T)$ /* Markov Blanket of $T$ */

1. Initialize: $MB(T) \emptyset$
2. Repeat
3. $Y = \arg \max_{F \in (U \cup MB(T) \backslash \{T\})} \text{dep}(T, F|MB(T))$
4. If $T \perp Y|MB(T)$ then
5. $MB(T) = MB(T) \cup \{Y\}$
6. Until $MB(T)$ does not change
7. For each $F \in MB(T)$ do
8. If $T \perp F|MB(T) \backslash \{F\}$ then
9. $MB(T) = MB(T) \backslash \{F\}$
10. Return $MB(T)$

---

To effectively increase the efficiency and accuracy of feature extraction, an improved Markov Blanket (IAMB) algorithm was developed and is described as follows.

In Algorithm 2, The IAMB algorithm consists of two phases: lines 2–9 denote the growing phase and lines 10–14 denote the shrinking phase. The main work of the growing phase is to find a Markov boundary of the target $T$, and all of the related factors will be redundant to $T$ in this boundary. The shrinking phase of IAMB tries to further eliminate the redundancy in the Markov boundary discovered in the prior phase. Note that the conditional independence test in IAMB is
controlled by the preselected threshold $\epsilon$, which will be nominally set to 0.01, as suggested in Aliferis et al.\textsuperscript{36} Note that $I(\cdot | \cdot)$ in the IAMB algorithm represents the conditional mutual information, where the size of the conditioning set in $I(\cdot | \cdot)$ will change at each iteration. In this study, the conditional mutual information that was implemented was put forward by Cooper et al.\textsuperscript{37}

The evaluation method

Two commonly used traffic injury severity classification algorithms named BN\textsuperscript{24} and CART 4.5(C4.5)\textsuperscript{26} were employed to generate classification error rates (CERs), and the characteristics of each algorithm are described as follows:

- **BN.** This is an annotated directed cyclic graph; the basic network is connected by a node and the node of the arrow or the directed edge. The node represents the random variable that would be affected by the events. The arrow or directed edge is equal to the relationship between the factors and events. The directed edge from the parent to children nodes and the contingent probability express the degree of influence between the parent and children nodes. The K2 search algorithm was employed to perform the BN structure in this study.

- **C4.5.** The C4.5 analysis is an effective algorithm to conduct prediction problems. The development of C4.5 model consists of four steps. The first step is data preprocessing; the data are always divided into two subsets (training set and testing set) and the target variable should be discrete. The second step is the calculation of Information Gain of each attribute—the calculation method is similar to the algorithm ID3. The next step is tree growing; according to step 2, the decision tree generation is carried out while the

---

Algorithm 2. IAMB algorithm

**Input:** Training dataset $D$ ($F, T$) /* $F$ is the factor set and $T$ is the target */

**Output:** $MB(T)$ /* Markov Blanket of $T$ */

1 **Initialize:** $MB(T) = \emptyset$
2 **Repeat**
3 **For each** $F \in F - MB(T)$ do
4 Find $F_{max}$ satisfying $\max_F I(F; T | MB(T))$
5 **If** $I(F_{max}; T | MB(T)) \geq \epsilon$
6 **Then** $MB(T) \leftarrow F_{max}$
7 **End**
8 **End**
9 **Until** $MB(T)$ does not change
10 **For each** $F \in MB(T)$ do
11 **If** $I(F; T | MB(T) - \{F\}) \leq \epsilon$
12 **Then** $MB(T) \leftarrow MB(T) - \{F\}$
13 **End**
observed values of each attribute are equal to the division subsets. The last step is the new dataset classified based on the classification rules.

The calculation process is carried out in WEKA, and all of two classification algorithms had already been integrated in this platform. The data analysis experiment is conducted using a 64-bit Windows computer with 2.4 GHz CPU, 8 GB RAM. The CER and accuracy are computed as follows

\[ CER = \frac{FP}{FP + TN} \]  \hspace{1cm} (1)

\[ Accuracy = \frac{TP}{FP + TP} \]  \hspace{1cm} (2)

where \( FP \) represents the number of events incorrectly classified as belonging to class A, and \( TN \) represents the number of events correctly classified as not belonging to class A. \( CER \) could represent the percentage of members of class A incorrectly classified as belonging to class A. \( TP \) represents the number of events correctly classified as belonging to class A.

The data

Description of FARS data

The FARS is a comprehensive database that became operational by NHTSA in 1975, and a majority of fatal traffic crashes are covered in this system in the United States. The source of FARS data is from police crash reports (PARS), and about 100 data elements for a crash that includes information from the driver, vehicle condition, and crash level are collected.

In order to identify the common features associated with the occurrence of injury fatality, the data of fatal crashes for 5 years in 50 US states were used in this study. Feature selection methods and machine learning algorithms were employed to conduct the datasets. However, data quality is a quite important issue, while the result of machine learning is determined by these issues. Therefore, the technology of data mining and preprocessing is essential to analyze the cause of traffic injury fatality.

Data preprocessing

Considering the accuracy of analysis findings between traffic injury fatality and impact factors, the FARS data need to be cleaned before use. The proposed data preprocessing method for analyzing FARS data consists of six filtering rules:

- The injury severity level was divided into three levels (0 = no injury, 1 = injury, and 2 = fatality) according to the worst-injured occupant. As an example, the possible injury and suspected minor injury in FARS are
classified as an injury crash. The unknown values (such as 5 = uninjured; severity unknown) are deleted;
- Remove the variables while all the values of the variable are 0 or otherwise;
- Remove the records while the value is no-numeric or missing;
- Remove the records with negative values (such as the value is $-1$; blank in driver-related factor);
- Remove outlying or low values;
- Remove or replace the variable that it had already confirmed with no-correlation with injury severity in previous study or expert evaluation.

After these filters, the number of selected crashes records was 45,490, 39,953, 40,095, 40,095, 37,427, respectively, from 2010 to 2014, and the detailed description of the datasets is presented in Figure 2. Moreover, 29 variables (e.g. atmospheric condition, day of crash, and so on) and one target variable were coded based on six above filtering rules, and the general information of the selected variables is shown in Table 1.

**Results and discussion**

**The result of feature selection**

In order to excavate the main factors associated with traffic injury severity, six datasets in FARS from 2010 to 2014 are applied in our study, and all of the datasets had already completed data preprocessing before our experiments. General information of six datasets is presented in Table 2, and the minimum descriptive length (MDL) method is adopted to discrete continuous features while the continuous and mixed datasets existed.

To evaluate the performance of IAMB in training speed in processing large sample datasets, three widely used feature selection algorithms—FCBF (fast correlation–based feature selection), ReliefF, and Wrapper-SVM (support vector machine)—are employed as comparisons (see Figure 3). The brief introduction about three contrast algorithms is as follows.
FCBF (fast correlation–based feature selection). The measurement criterion is symmetrical uncertainty (SU) in this algorithm. First, all the features are ranked based on the correlation test, then the redundant features are eliminated using approximate Markov Blankets, and the judgment criteria are as follows: if $SU(F_1; C) > SU(F_2; C)$ and $SU(F_1; C) > SU(F_1; F_2)$, then $F_2$ is the redundant feature of $F_1$.

Table 1. Variable description.

| ID | Variable                              | Type               | Description                                                                 |
|----|---------------------------------------|--------------------|-----------------------------------------------------------------------------|
| 1  | Atmospheric condition                  | Qualitative        | Atmospheric type (0–11)                                                    |
| 2  | Day                                   | Qualitative        | Day of crash (1–31)                                                        |
| 3  | Hour                                  | Qualitative        | Time of crash (0–23)                                                       |
| 4  | Month                                 | Qualitative        | Month of crash (1–12)                                                      |
| 5  | Crash-related factors (1)             | Qualitative        | Reason caused crash (0–28)                                                 |
| 6  | Day of week                           | Qualitative        | 1, Monday; 2, Tuesday; 3, Wednesday; 4, Thursday; 5, Friday; 6, Saturday; |
|    |                                       |                    | 7, Sunday                                                                   |
| 7  | Large truck                           | Qualitative        | 0, Not heavy truck related; 1, Heavy truck related                          |
| 8  | Light condition                       | Qualitative        | 1, Daylight; 2, Dark-not lighted; 3, Dawn; 4, Dusk                         |
| 9  | Manner of collision                   | Qualitative        | The collision type (0–11)                                                  |
| 10 | National highway system               | Qualitative        | 0, No; 1, Yes                                                              |
| 11 | Relation to junction                  | Qualitative        | Specific location (1–20)                                                   |
| 12 | Speeding                              | Qualitative        | 0, Not speeding; 1, Speeding                                               |
| 13 | Type of intersection                  | Qualitative        | Type of intersection (1–10)                                                |
| 14 | Work zone                             | Qualitative        | 0, None; 1, Construction; 2, Maintenance; 3, Utility; 4, Work zone         |
| 15 | Age                                   | Continuous         | Age of the most severity injury person                                     |
| 16 | Alcohol test result                   | Continuous         | Alcohol percentage                                                         |
| 17 | Person type                           | Qualitative        | 1, Driver of a motor vehicle in-transport; 2, Passenger of a motor vehicle |
|    |                                       |                    | in-transport                                                               |
| 18 | Drug involvement                      | Qualitative        | 0, No; 1, Yes                                                              |
| 19 | Sex                                   | Qualitative        | 1, Male; 2, Female                                                          |
| 20 | Driver alcohol involvement            | Qualitative        | 0, No; 1, Yes                                                              |
| 21 | Driver-related factor (1)             | Qualitative        | Driver relation factor to crash (0–92)                                     |
| 22 | License state                         | Qualitative        | The name of state (0–56)                                                   |
| 23 | Crash type                            | Qualitative        | The type of crash (0–98)                                                   |
| 24 | Driver distracted                     | Qualitative        | The type of distraction (0–19)                                             |
| 25 | Driver vision obscured by             | Qualitative        | The type of obstruction (0–14)                                             |
| 26 | Roadway alignment                     | Qualitative        | 0, No traffic way; 1, Straight; 2, Curve right; 3, Curve left              |
| 27 | Roadway surface condition             | Qualitative        | 0, Non-traffic way; 1, Dry; 2, Wet; 3, Snow; …; 11, Mud, dirt, or gravel  |
| 28 | Roadway surface type                  | Qualitative        | 0, Non-traffic way; 1, Concrete; 2, Blacktop; 3, Brick or Block; 4, Slag; |
|    |                                       |                    | 5, Dirt                                                                    |
| 29 | Speed limit                           | Continuous         | Posted speed limit                                                         |
| 30 | Injury level                          | Qualitative        | Target variable (0, No-injury; 1, Injury; 2, Fatality)                     |
ReliefF. It is a well-known feature sorting method based on distance. The feature selection criterion of this method is to select the feature with the maximum distinguishing distance of different class tags and the minimum distinguishing distance of the same class tags. The algorithm needs to set the nearest neighbor number $k$ and the number of participating samples $m$ before training, and $k = 5$, $m = 30$ in this study.

Wrapper-SVM. The characteristics of this algorithm are that the SVM algorithm is embedded into the Wrapper class attribute sorting algorithm, and the optimal or locally optimal attribute sets are obtained based on the classification accuracy.

| Years     | Samples  | Features | Type    | Classes |
|-----------|----------|----------|---------|---------|
| 2010      | 45,490   | 29       | Numeric | 3       |
| 2011      | 39,953   | 29       | Numeric | 3       |
| 2012      | 40,095   | 29       | Numeric | 3       |
| 2013      | 40,095   | 29       | Numeric | 3       |
| 2014      | 37,427   | 29       | Numeric | 3       |
| 2010–2014 | 203,060  | 29       | Numeric | 3       |

Figure 3. Comparison with different feature selection algorithms.
A comparison between IAMB and four other feature selection algorithms with respect to the number of selected features is shown in Figure 3.

The results shown in Figure 3 clearly present that the dimension of features is decreased after using five feature selection algorithms, and the IAMB algorithm is superior to other four algorithms in different four datasets (the number of selected features are 9, 12, 7, and 7, respectively). It ranked second in the other two datasets (the number of selected features is 7 and 7, respectively). Meanwhile, training time is also considered as an essential indicator to evaluate the performance of a feature selection algorithm; thus, the training time is also recorded and can be seen in Table 3.

As shown in Table 3, the training time in six datasets shows that IAMB (1.85, 1.37, 1.44, 1.32, 1.00, and 4.21 s, respectively) outperforms the other algorithms. Although the Wrapper-SVM extracted less number of selected features in 2011 and 2013 (the number of selected features is 6 and 6, respectively) in Figure 3, the training time of Wrapper-SVM is more than 600 s. Meanwhile, the results in Figure 4(a) and (b) show that the IAMB algorithm presented the best performance irrespective

![Figure 4](image-url)

**Table 3.** The training time of different feature selection algorithms.

| Datasets     | MB  | FCBF | ReliefF | IAMB | Wrapper-SVM |
|--------------|-----|------|---------|------|-------------|
| 2010         | 1.98| 1.89 | 2.4     | 1.85*| >600        |
| 2011         | 1.41| 1.55 | 2.21    | 1.37*| >600        |
| 2012         | 1.52| 1.51 | 2.02    | 1.44*| >600        |
| 2013         | 1.36| 1.47 | 2.08    | 1.32*| >600        |
| 2014         | 1.08| 1.36 | 1.79    | 1.00*| >600        |
| 2010–2014    | 4.41| 4.08 | 5.31    | 4.21*| >600        |

MB: Markov Blanket; FCBF: fast correlation–based filter; IAMB: improved Markov Blanket; SVM: support vector machine.

*Bold values represent the highlighted number results of IAMB approach in this research, the details of significance have been analyzed in paper.
of the number of selected features and the training speed in most of the datasets, which indicated that the IAMB selected fewer features using the shortest training time.

In addition, the results of using IAMB presented in Table 4 indicate that atmospheric condition, hour of crash, alcohol test result, crash type, and the driver’s distraction are critical factors that influenced traffic injury severity among different years from 2010 to 2014. There also exist many features that are not common in different datasets, such as work zone, month of crash, and so on.

### Calibration and validation of feature selection

In order to check the quality of the selected features by IAMB, the Pearson correlation coefficient test and the CER test were employed. The relationship between the selected features and injury severity was also analyzed, and the different classification algorithms were employed to classify different injury severity levels based on the selected features.

**The result of Pearson correlation coefficient test.** As shown in Figure 5, the results present that the vast majority of selected features are significantly correlated with the traffic injury severity ($p < 0.01$) in different datasets, where $p$ value $< 0.01$ indicates the statistical significance of the correlation between the two selected variables. Only driver’s distraction (ID: 24) in 2010 and 2011, driver alcohol involvement (ID: 20) in 2010, month of crash (ID: 4) in 2010–2014, and license state (ID: 22) are not significantly correlated with injury severity. Thus, the result of correlation test indicates that the feature selection by IAMB is effective.

In addition, Figure 5 also records the calculation results of the correlation coefficient between selected features and injury severity. The absolute value of correlation coefficient of alcohol test result (0.325 in 2010, 0.447 in 2011, 0.398 in 2012, 0.441 in 2013, 0.482 in 2014, and 0.268 in 2010–2014) is greatest in different datasets, which indicates that the alcohol test result is a critical impact factor that caused the occurrence of traffic injury. Meanwhile, the features of 1, 3, 23, and 24 are also associated with traffic injury severity.

| Years     | Selected features using IAMB |
|-----------|------------------------------|
| 2010      | 1, 3, 4, 16, 17, 18, 20, 23, 24 |
| 2011      | 1, 3, 16, 17, 18, 23, 24      |
| 2012      | 1, 3, 4, 12, 16, 17, 18, 19, 20, 22, 23, 24 |
| 2013      | 1, 3, 16, 17, 20, 23, 24      |
| 2014      | 1, 3, 16, 17, 23, 24          |
| 2010–2014 | 1, 3, 4, 16, 18, 23, 24       |
| Common features | 1, 3, 16, 23, 24 |

IAMB: improved Markov Blanket.
The result of CER test. Two famous and most frequently used classifiers, BN and C4.5, in traffic accident analysis are adopted to generate the accuracy of

Figure 5. Correlation test between injury severity and selected features in different years: (a) in 2010, (b) in 2011, (c) in 2012, (d) in 2013, (e) in 2014, and (f) in 2010–2014.
classification and CER on six datasets with selected features by different feature selection algorithms. Three types of selected features extracted by genetic search, rank search, and un-selected were adopted as a comparison. Ten-fold cross-validation method was applied to evaluate the accuracy of different algorithms.

The results of Figures 6 and 7 show that the accuracy of classification with IAMB in BN and C4.5 is better than the results with compared algorithms and un-selected in most of the datasets. Although the improvement of accuracy of classification is not significantly greater, the huge losses to humans and the society due to fatality injury are worthy of consideration. In addition, the result of Figure 6 also indicated that a better classification results could be achieved by C4.5 in FARS data analysis, and the best accuracy of classification appeared in 2014 by C4.5 reached 68.71%.

For further analysis, the CER is calculated to visualize the difference of IAMB compared with the other selected feature selection algorithms under two classifiers on six datasets. Figure 8(a) and (b) presents that the value of CER of IAMB under BN and C4.5 is lower when compared with that in most of the datasets, respectively. The lowest CER of IAMB appeared in 2014 under BN and reached 16.04%. The result of CER indicates that IAMB achieves better performance in most of the datasets, and it is useful to extract the critical factors that influenced traffic injury severity in FARS.
Discussions on common and significant impact factors

As shown in Table 4, when using the IAMB algorithm, five variables were identified as significant and common factors that contribute to traffic injury severity. They are as follows: (1) atmospheric condition, (2) time of crash, (3) alcohol test result, (4) crash type, and (5) driver’s distraction. The values of different factors were defined by combining requirements from the manual of FARS and

Figure 7. The accuracy of C4.5.

Figure 8. The CER of (a) BN and (b) C4.5 in different feature selection algorithms.

Discussions on common and significant impact factors

As shown in Table 4, when using the IAMB algorithm, five variables were identified as significant and common factors that contribute to traffic injury severity. They are as follows: (1) atmospheric condition, (2) time of crash, (3) alcohol test result, (4) crash type, and (5) driver’s distraction. The values of different factors were defined by combining requirements from the manual of FARS and
engineering experiences in previous studies. As an example, the alcohol test result was divided into three states: it was defined as drunk driving while the alcohol test result is greater than 0.08% blood alcohol concentration (BAC) and the state is defined as others while the value of alcohol test is greater than 0.94% BAC. Moreover, in order to deeply analyze the relationship between the common features and traffic injury severity, the crash rate (CR) was defined and was calculated as follows

\[ CR = \frac{CN_1 = k}{CN_1 + CN_2 + CN_3} \quad (k = 1, 2, 3) \]

where \( i \) is the mean level of traffic injury severity, and \( CN \) is the the number of crash.

The description of the selected variables, together with their \( CN \) and CR, is presented in Table 5. The statistical analysis indicates that the value of \( CN \) and CR of different levels of traffic injury severity in different states of selected features is different. Therefore, the detailed interpretation for these five features is needed and is provided in Table 5.

**Atmosphere condition.** As shown in Table 5, compared with crash occurrences during fine weather condition, the rate of fatality crashes decreases 1% in rain day and increases 1.9% in fog or worse. This finding is consistent with Edwards\(^{43}\) who found that accident severity in rain decreases significantly when compared with fine weather. The reason for the results may be high speed and paying less attention to driving during fine weather. Moreover, the rate of no-injury crashes in fine weather is the highest when compared with other weather conditions, which indicates that although the crash easily occurs in fine weather, the damage of an accident is lower than in other weather conditions.

**Time of crash.** The classification of time of crash occurrence was similar to that of Huang et al.'s\(^{23}\) who defined the time of crash as day time (10:00 a.m.–5:00 p.m.), peak time (7:00 a.m.–10:00 a.m. and 5:00 p.m.–8:00 p.m.), and night time (8:00 p.m.–7:00 a.m.). In previous studies,\(^{2,23,24}\) it had already been confirmed that more severe injuries occur during darkness, and crashes occurring in peak time are associated with lower severity when compared with those in day time and night time. The visibility is poorer at night time, and the drivers’ attention is more easily to distract. This is the main cause of traffic accidents at night. In contrast, at peak time, the driving speed will be lower and dangerous driving behaviors such as forced lane changes will be reduced, leading to lower severity of accidents. This coincides with the results found in this study.

**Alcohol test result.** The alcohol test result was classified into three types: normal driving (0%–0.08% BAC), drunk driving (0.09%–0.94% BAC), and others (higher than 0.94% BAC). The results shown in Table 5 present that 62% of the occurrence of crashes are fatal accidents while drivers are convicted of drunk driving, and the rate of fatal crashes is significantly higher than in normal driving. Previous studies have observed that drunk driving impairs driving abilities and large numbers of
| Common features          | Values        | No-injury | Injury | Fatality | Total |
|--------------------------|---------------|----------|--------|----------|-------|
|                          |               | CN       | CR (%) | CN       | CR (%) | CN       | CR (%) | CN       | CR (%) |
| Atmospheric condition    | Normal        | 39,048   | 25.7   | 53,109   | 35.0   | 59,750   | 39.3    | 15,197   |        |
|                          | Rain          | 3383     | 24.1   | 5275     | 37.6   | 5390     | 38.3    | 14,048   |        |
|                          | Sleet, hail   | 156      | 15.9   | 458      | 45.8   | 365      | 37.3    | 979      |        |
|                          | Snow          | 961      | 22.8   | 1609     | 38.3   | 1633     | 38.9    | 4203     |        |
|                          | Fog or worse  | 7685     | 24.0   | 11,573   | 36.3   | 12,665   | 39.7    | 31,923   |        |
| Time of crash            | Day time      | 19,105   | 24.4   | 28,877   | 36.9   | 30,286   | 38.7    | 78,268   |        |
|                          | Night time    | 18,040   | 24.7   | 25,049   | 34.4   | 29,815   | 40.9    | 72,904   |        |
|                          | Peak time     | 14,421   | 27.2   | 18,493   | 34.9   | 20,114   | 37.9    | 53,028   |        |
| Alcohol test result      | Normal        | 11,817   | 20.8   | 13,924   | 24.5   | 31,028   | 54.7    | 56,769   |        |
|                          | Drunk driving | 2967     | 13.9   | 5059     | 23.6   | 13,377   | 62.0    | 21,403   |        |
|                          | Others        | 36,782   | 29.2   | 53,436   | 42.4   | 35,810   | 28.4    | 126,028  |        |
| Crash type               | Category 1    | 20,317   | 25.9   | 22,883   | 29.2   | 35,127   | 44.8    | 78,327   |        |
|                          | Category 2    | 6675     | 29.7   | 8084     | 36.0   | 7703     | 34.3    | 22,462   |        |
|                          | Category 3    | 6987     | 17.8   | 17,350   | 44.2   | 14,924   | 38.0    | 39,261   |        |
|                          | Category 4    | 5975     | 26.7   | 8100     | 36.3   | 8228     | 36.9    | 22,303   |        |
|                          | Category 5    | 4168     | 21.3   | 8242     | 42.1   | 7189     | 36.7    | 19,599   |        |
|                          | Category 6    | 7444     | 33.5   | 7760     | 34.9   | 7044     | 31.7    | 22,248   |        |
| Driver’s distraction     | Not distracted| 41,199   | 25.5   | 56,980   | 35.2   | 63,530   | 39.3    | 161,709  |        |
|                          | Distracted    | 10,504   | 24.0   | 15,613   | 35.7   | 17,657   | 40.3    | 43,774   |        |

Category 1: single driver; category 2: same traffic way, same direction; category 3: same traffic way, opposite direction; category 4: changing traffic way, vehicle turning; category 5: intersecting paths; category 6: miscellaneous; CN: number of crash; CR: crash rate.
fatal and injury crashes were related to drunk driving.\textsuperscript{44} Stübiger et al.\textsuperscript{45} found that drunk driving increases the probability of injuries and fatal injury in a crash.

\textbf{Crash type.} The results shown in Table 5 indicate that the crash type of single driver was found to be more involved in fatality accidents, with the fatality rate reached at 44.8%. The highest CR of no-injury appeared in category 2, which indicated that a same traffic way and same direction accident resulting in lower accident severity. De Oña et al.\textsuperscript{46} found that the probability of killed or seriously injured accidents increased while the rollover (belong to category 1 in this study) occurred.

\textbf{Driver’s distraction.} It was found that the crash rate of injury and fatality (the CR reached 35.7% and 40.3%, respectively) was increased when the driver is distracted when compared with a non-distracted driver. This is consistent with Neyens and Boyle\textsuperscript{47} who found that the passengers are easier to suffer from severe injuries when the driver is distracted than with a non-distracted driver. In addition, the results shown in Table 5 also indicate that crashes that occur always result in no-injury severity while the driver is non-distracted or attentive.

\section*{Conclusion and recommendations}

A feature selection (IAMB) algorithm was proposed to extract the main factors that are associated with the injury severity while traffic accident occurred. It is useful to help the department of transportation to develop the related policy to decrease the occurrence of severe injuries. This algorithm only needed simple preprocessing about the initial data, and it is helpful to analyze the big data. The FARS data from 2010 to 2014 were preprocessed and applied as the datasets for the analysis. Two commonly used classification algorithms (BN and C4.5) and three indicators (correlation coefficient, accuracy of classification, and CER) have been employed to examine the performance and effectiveness of IAMB algorithm. The results of validation and verification indicate that IAMB is a good feature selection algorithm to analyze injury severity of FARS data, with significant correlation between the selected factors and injury severity, with the increase in the accuracy of classification, and with the decrease in the classification rate.

The IAMB algorithm selected five impact factors (including atmosphere condition, time of crash, alcohol test result, crash type, and driver’s distraction) that significantly influenced traffic injury severity from the initial 29 variables. These five factors are regarded as common features which were selected in all six dataset (the FARS data from 2010 to 2014). It was found that crashes occurred in bad weather condition (e.g. fog or worse), in night time, in drunk driving, in crash type of single driver, and in distracted driving, which are associated with more severe injuries. Specially, results indicated that crashes in peak time resulted in low severity. The result of the Pearson correlation coefficient test shown in Figure 4 indicates that the alcohol test result is the most important factor that significantly influenced traffic injury severity. All in all, the results are consistent with the previous research.

This study shows that the combined use of feature selection (IAMB) algorithm and data analysis method provides new insights and ideas on the extraction of main
causes of crash severity for FARS data. It is worthwhile to do the feature selection before applying other machine learning techniques to predict traffic injury severity while a traffic accident occurred.

However, no study is without limitations. In this study, due to the incomplete and bias of FARS data, the best of accuracy of injury severity classification only reached 68.71%. A better dataset is needed to achieve a high enough traffic injury severity prediction model. Different results might have obtained if other types of data had been applied. Moreover, in order to increase the efficiency and simple data processing, the traffic injury severity was concluded into three states (no-injury, injury, and fatality) and the unknown value of all the variables was deleted. It might be influenced by the final result of crash data analysis.

Data availability
FARS is a nationwide census providing NHTSA, Congress and the American public yearly data regarding fatal injuries suffered in motor vehicle traffic crashes. The data can be accessed through the link https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is sponsored by the National Natural Science Foundation of China (51805169, 51605350, and U1764262); Wuhan Science and Technology Project (2019010702011301); and the National Engineering Laboratory for Transportation Safety & Emergency Informatics in China (YW170301-09). This work was also supported by the China Scholarship Council (CSC).

ORCID iD
Yi He https://orcid.org/0000-0002-3005-6772

References
1. Goetzke F and Islam S. Determinants of seat belt use: a regression analysis with FARS data corrected for self-selection. *J Safety Res* 2015; 55: 7–12.
2. De Oña J, López G, Mujalli R, et al. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. *Accid Anal Prev* 2013; 51: 1–10.
3. Tang J, Liang J, Han C, et al. Crash injury severity analysis using a two-layer Stacking framework. *Accid Anal Prev* 2019; 122: 226–238.
4. Karrer K and Matthias R. *Effects of driver fatigue monition: An expert survey*. Berlin; Heidelberg: Springer, 2007, pp. 324–330.
5. Ma Z, Zhao W, Chien SIJ, et al. Exploring factors contributing to crash injury severity on rural two-lane highways. *J Safety Res* 2015; 55: 171–176.

6. Santamarina-Rubio E, Perez K, Olabarria M, et al. Gender differences in road traffic injury rate using time travelled as a measure of exposure. *Accid Anal Prev* 2014; 65: 1–7.

7. Hanrahan RB, Layde PM, Zhu S, et al. The association of driver age with traffic injury severity in Wisconsin. *Traffic Inj Prev* 2009; 10(4): 361–367.

8. Paleti R, Eluru N and Bhat CR. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accid Anal Prev* 2010; 42(6): 1839–1854.

9. Michalaki P, Quddus MA, Pitfield D, et al. Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model. *J Safety Res* 2015; 55: 89–97.

10. Abu-Zidan FM and Eid HO. Factors affecting injury severity of vehicle occupants following road traffic collisions. *Injury* 2015; 46(1): 136–141.

11. Kononov J, Reeves D, Durso C, et al. Relationship between freeway flow parameters and safety and its implication for adding lanes. *Transp Res Rec* 2012; 2279: 118–123.

12. Boufous S, Finch C, Hayen A, et al. The impact of environmental, vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash. *J Safety Res* 2008; 39(1): 65–72.

13. Yu R and Abdel-Aty M. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Saf Sci* 2014; 63: 50–56.

14. Christoforou Z, Cohen S and Karlaftis MG. Vehicle occupant injury severity on highways: an empirical investigation. *Accid Anal Prev* 2010; 42(6): 1606–1620.

15. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans Neural Netw* 1994; 5(4): 537–550.

16. Fleuret F. Fast binary feature selection with conditional mutual information. *J Mach Learn Res* 2004; 5: 1531–1555.

17. Yang S. On feature selection for traffic congestion prediction. *Transp Res Part C Emerg Technol* 2013; 26: 160–169.

18. Zhang H, Lu G, Qassrawi MT, et al. Feature selection for optimizing traffic classification. *Comput Commun* 2012; 35(12): 1457–1471.

19. Li R, Pu S, Lu J, et al. A traffic flow status recognition method by combining fuzzy logic and rough set theory. In: *Eighth international conference of chinese logistics and transportation professionals*, 2009, pp. 3326–3332, https://ascelibrary.org/doi/10.1061/40996%28330%29488

20. Wang Y, Mabu S, Meng Q, et al. Multiple ODs routing algorithm for traffic systems using GA. In: *2010 IEEE world congress on computational intelligence (WCCI 2010)—2010 IEEE congress on evolutionary computation (CEC 2010)*, https://ieeexplore.ieee.org/document/5586424

21. Biesiada J and Duch W. Feature selection for high-dimensional data: a Kolmogorov-Smirnov correlation-based filter. *Comput Recognit Syst* 2005; 30: 95–103.

22. Zhen L and Qiong L. A new feature selection method for internet traffic classification using ML. *Phys Procedia* 2012; 33: 1338–1345.

23. Huang H, Chin HC and Haque MM. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accid Anal Prev* 2008; 40(1): 45–54.

24. Yu R and Abdel-Aty M. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accid Anal Prev* 2014; 62: 161–167.
25. Chang LY and Wang HW. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 2006; 38(5): 1019–1027.
26. Kashani AT and Mohaymany AS. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf Sci 2011; 49(10): 1314–1320.
27. Abdelwahab HT and Abdel-aty M. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp Res Rec 1997; 1746: 6–13.
28. Chengwu J, Guiyan J, Ximin L, et al. Dangerous situation recognition method of driver assistance system. In: 2006 IEEE International Conference on Vehicular Electronics and Safety, 2006, pp. 169–173, https://ieeexplore.ieee.org/document/4234012
29. Yan L, Huang Z, Zhang Y, et al. Driving risk status prediction using Bayesian networks and logistic regression. IET Intell Transp Syst 2017; 11(7): 431–439.
30. Zou Y, Zhong X, Tang J, et al. A copula-based approach for accommodating the underreporting effect in wildlife-vehicle crash analysis. Sustain 2019; 11(2): 13.
31. Koller D and Sahami M. Toward optimal feature selection. In: International conference on machine learning, 1996, pp. 284–292, https://dl.acm.org/citation.cfm?id=3091731
32. Yan L, Zhang Y, He Y, et al. Hazardous traffic event detection using Markov blanket and sequential minimal optimization (MB-SMO). Sensors (Basel) 2016; 16(7): E1084.
33. Tsamardinos I, Aliferis C, Statnikov A, et al. Algorithms for large scale Markov blanket discovery. In: FLAIRS Conference, 2003, pp. 376–381, https://www.aaai.org/Papers/FLAIRS/2003/Flairs03-073.pdf
34. Zhang Y and Zhang A. Feature subset selection with cumulative conditional mutual information minimization. Expert Syst Appl 2012; 39(5): 6078–6088.
35. Zhang Y, Zhang Z, Liu K, et al. An improved IAMB algorithm for Markov blanket discovery. J Comput 2010; 5(11): 1755–1761.
36. Aliferis C, Statnikov A and Tsamardinos I. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. J Mach 2010; 11: 235–284.
37. Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 1997; 9(2): 107–138.
38. Mango NK and Garthe E. Why people die in motor vehicle crashes: linking detailed causes of death with FARS data. System 1998; 412: 0148–7191.
39. Tseng WS, Nguyen H, Liebowitz J, et al. Distractions and motor vehicle accidents: data mining application on fatality analysis reporting system (FARS) data files. Ind Manag Data Syst 2005; 105(9): 1188–1205.
40. Egea S, Rego Manez A and Carro B. Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments. IEEE Internet Things J 2018; 3(5): 1616–1624.
41. Chen Z, Wu C and Huang Z. Dangerous driving behavior detection using video-extracted vehicle trajectory histograms. J Intell Transp Syst 2017; 21(5): 409–421.
42. Yan L, Wu C, Zhu D, et al. Driving mode decision making for intelligent vehicles in stressful traffic events. Transp Res Rec J Transp Res Board 2017; 2625(1): 9–19.
43. Edwards JB. The relationship between road accident severity and recorded weather. J Safety Res 1998; 29(4): 249–262.
44. Babor TF, Caetano R, Casswell S, et al. Alcohol: No ordinary commodity: Research and public policy. Oxford: Oxford University Press, 2010.
45. Stübig T, Petri M, Zeckey C, et al. Alcohol intoxication in road traffic accidents leads to higher impact speed difference, higher ISS and MAIS, and higher preclinical mortality. *Alcohol* 2012; 46(7): 681–686.

46. De Ona J, Mujalli RO and Calvo FJ. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accid Anal Prev* 2011; 43(1): 402–411.

47. Neyens DM and Boyle LN. The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers. *Accid Anal Prev* 2008; 40(1): 254–259.

**Author biographies**

Lixin Yan received his Ph.D. degree in Transportation Engineering from Wuhan University of Technology in 2017. He has been a visiting scholar at University of Wisconsin-Madison, USA, during 2015-2016. He is currently an Associate Professor at College of Transportation and Logistics, East China Jiaotong University, Jiangxi, China. His current research interests include traffic safety, traffic management and control, and intelligent transportation systems.

Yi He is currently an Assistant Professor in Wuhan University of Technology, China. He received M.S. (2012) and Ph.D (2016) degrees in intelligent transportation system engineering from the Wuhan University of Technology, China. He was also a co-trained Ph.D at California PATH, University of California, Berkeley (2014). He worked as a Post-doc in California PATH, University of California, Berkeley, CA, USA. (2017). His research interests include vehicle system modeling, micro/macro simulation, autonomous vehicle safety and human factors.

Ms. Lingqiao Qin is a Research Assistant at the University of Wisconsin-Madison. She received the B.S. in Transportation Engineering from the Beijing Jiaotong University, Beijing, China, the M.S. in Transportation Safety Engineering from the George Washington University, Washington, D.C., US, and the M.S. in Industrial and Systems Engineering from University of Wisconsin-Madison, WI, US. Lingqiao is currently working toward her Ph.D. degree in Transportation Engineering at the University of Wisconsin-Madison, WI, US. Her research interests include traffic operations, the next generation of transportation (autonomous and connected vehicles), and using advanced technologies such as driving simulator and eye trackers to improve the design, operations, and safety of all elements in transportation. Lingqiao has published many peer-reviewed journal papers and has given several presentations at national and international conferences.

Chaozhong Wu is currently a Professor in Intelligent Transport System Research Center, Wuhan University of Technology, China. He received his M.S. and Ph.D. degrees from Wuhan University of Technology in 1999 and 2002, respectively. He is Membership of National ITS Standardization committee, China, and Associate Secretary-General of the Committee of Youth Scientific and Technological Workers for Transportation, China. He was a visiting scholar of University of Regina, Canada. His research interests are in traffic safety, driver behavior, intelligent transportation systems, and intelligent vehicle.

Dunyao Zhu received his B.E. degree and the M.S. degree in Wuhan University, Hubei, China, in 1983 and 1986, and the Ph.D degree in the University of Tokyo, Tokyo, Japan, in 1996; He is a Professor with the Department of Intelligent Transport Systems Research
Center, Wuhan University of Technology, Wuhan, China. His research has focused on five major areas: 1) intelligent transportation system; 2) connected vehicle, and 5) autonomous driving.

Bin Ran received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 1986; the M.S. degree in civil engineering from the University of Tokyo, Tokyo, Japan, in 1989; and the Ph.D. degree in civil engineering from the University of Illinois, Chicago, IL, USA, in 1993. He is a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin -Madison, Madison, WI, USA. He has led the development and deployment of various traffic information systems and technologies in the United States and China. He is the author of two leading textbooks on dynamic traffic networks. He has coauthored more than 100 journal papers and more than 200 referenced papers at national and international conferences. His research has focused on five major areas: 1) intelligent transportation system; 2) dynamic transportation network and traffic modeling; 3) development of cellular probe technologies; 4) connected vehicle, and 5) Internet of mobility.