Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 1257627, 13 pages
https://doi.org/10.1155/2020/1257627

Research Article

An Alternative Method for Traffic Accident Severity Prediction: Using Deep Forests Algorithm

Jing Gan,[1] **Linheng Li**,[1] **Dapeng Zhang**,[2] **Ziwei Yi**,[1] **and Qiaojun Xiang**[1]

[1] *Jiangsu Key Laboratory of Urban ITS, School of Transportation, Southeast University, Nanjing 211189, China*
[2] *Big Data Research Center, Southwestern University of Finance and Economics, Chengdu 611130, China*

Correspondence should be addressed to Qiaojun Xiang; xqj@seu.edu.cn

Received 10 February 2020; Revised 12 April 2020; Accepted 1 May 2020; Published 21 December 2020

Academic Editor: Qing Cai

Traffic safety has always been an important issue in sustainable transportation development, and the prediction of traffic accident severity remains a crucial challenging issue in the domain of traffic safety. A huge variety of forecasting models have been proposed to meet this challenge. These models gradually evolved from linear to nonlinear forms and from traditional statistical regression models to current popular machine learning models. Recently, a machine learning algorithm called Deep Forests based on the decision tree ensemble has aroused widespread concern, which was proposed for the first time by a research team of Nanjing University. This algorithm was proved to be more accurate and robust in comparison with other machine learning algorithms. Motivated by this benefit, this study employs the UK road safety dataset to propose a novel method for predicting the severity of traffic accidents based on the Deep Forests algorithm. To verify the superiority of our proposed method, several other machine learning algorithm-based perdition models were implemented to predict traffic accident severity with the same dataset, and the prediction results show that the Deep Forests algorithm present good stability, fewer hyper-parameters, and the highest accuracy under different level of training data volume. It is expected that the findings from this study would be helpful for the establishment or improvement of effective traffic safety system within a sustainable transportation system, which is of great significance for helping government managers to establish timely proactive strategies in traffic accident prevention and effectively improve road traffic safety.

1. Introduction

Traffic safety has always been an important issue in sustainable transportation development. Traffic accidents will have some negative impacts on society, including casualties, traffic jams, and environmental pollution, which are not conducive to the sustainable and healthy development of the transportation system. With the gradual improvement of the level of automated information systems, in recent years, some government agencies and transportation industry companies have been committed to the development of intelligent transportation systems to help the sustainable development of transportation. Traffic accident prediction is a crucial and challenging issue in the domain of intelligent traffic safety management system; it is of great significance for analyzing the future development trend of traffic accidents and implementing proactive prevention measures under existing road traffic conditions. To improve traffic safety management and control, it is necessary to seek timely and accurate methods for predicting traffic accident severity. In recent years, with the rapid development of science and technology, the advanced technology used in transportation has been strengthened at an unprecedented level. Unfortunately, these advanced technologies have no obvious advantages for the reduction of traffic accidents. Save LIVES-A road safety technical package 2017, issued by World Health Organization (WHO), indicated that road traffic accidents lead to the loss of over 1.2 million lives and cause nonfatal injuries to as many as 50 million people around the world each year, which are estimated to be the ninth leading cause of death across all age groups globally [1]. Road traffic crashes may be an everyday occurrence, but they are predictable and preventable. Therefore, every traffic researcher has the responsibility to think over the causes of traffic accidents and help the administration in solving the problem of reducing the probability of traffic accidents. Over the years, researchers have tried various traffic accident severity analysis models from different perspectives. These modeling analyses are to explore the relationship between accident severity and its influencing factors, among which the most widely used is the discrete selection model based on the Logit or Probit model (e.g., [2–6]). These studies have shown that accurate traffic accident severity prediction plays an important role in improving traffic safety management, because, based on accurate prediction, the prominent influencing factors in high-risk road sections could be found out to provide beneficial suggestions for improving road safety.

Latterly, with the advancement of computer science, the era of big data has come. Many scholars began to try to apply some intelligent classification models based on knowledge discovery for accident degree analysis modeling, such as the Bayesian model, neural network model, decision tree model, and random forest model [7–11]. All of these models have one common characteristic that they do not require any assumptions on the relationship between the independent variables and dependent ones. Mujalli et al. [7] used Bayesian networks to improve classifying the traffic accident, which results in a reduction in the misclassification of deaths and serious injuries. García de Soto et al. [8] found that Artificial Neural Networks (ANNs) can be used as a feasible method to predict the frequency of road traffic accidents. Zhang and Fan [9] presented a data mining model using ID3 and C4.5 decision tree algorithms to analyze the traffic collision data. Pu et al. [10] conducted Full Bayesian before-after analysis of safety effects (crash severity levels, crash types, and crash causes) of variable speed limit system based on crashes data. Dadashova et al. [5] estimated the impact of the influencing factors on road traffic accident severity through random forests. It is worth noting that, in the above methods, random forest is an integrated learning method for classification, regression, and other tasks, which is more accurate and robust than other existing algorithms and effective for large databases. Therefore, in recent years, this method has been widely applied to various traffic problems [12–16]. Liu and Wu [12] established a traffic congestion prediction model using the machine learning classification algorithm, random forest. Mudali [13] analyzed the traffic big data using two comparative parallel algorithms M5P rules and random forest regression from the regression model for determining the nature of traffic big data. Nadarajan et al. [14] predicted a probabilistic space-time representation of complex traffic scenarios by using random forest algorithms. Kwon and Park [16] analyzed the impact of weather factors on traffic safety levels using k-means clustering and random forest techniques, and the result showed that the proposed model outperforms the conventional traffic safety prediction models.

There are certainly some shortcomings in the random forest model. Some researchers try to continuously improve the RFs (random forests) even though it already has many advantages. Gao and Ke [17] employed a random survival forests model to analyze the incident duration analysis model and make a comparison with the traditional random forests model. The result shows that the random survival forests models are more accurate. Several researchers have proposed to incorporate RFs into the deep neural system [18–22]. The most representative of which is Deep Forests proposed by Zhou and Feng [18] in 2017. This algorithm with much fewer hyper-parameters was proved to achieve excellent performance in various domains by using the same parameter setting. Since this algorithm was recently proposed, there are almost no applications in the transportation field.

Road traffic accidents are the process of simultaneous damage to people or things caused by the miscoupling of dynamic and static factors (e.g., people, vehicles, roads, and the environment) [23–27]. The historical data of road traffic accidents can directly reflect the relationship between these factors during the accident. Benefitting from the excellent performance of Deep Forests, in this paper, we propose a traffic accident severity prediction method based on the Deep Forests algorithm, including data preprocessing, data feature selection, and accident severity perdition. After the data preprocessing is completed, we use the method of Random Forests to select the data features, which will be finally trained in Deep Forests algorithm. To the best of the authors' knowledge, this is the first time that the Deep Forests algorithm is used to predict the severity of traffic accidents. The correlations between each feature are inherently considered in the modeling. In addition, the final prediction results demonstrate that the proposed method for accident severity prediction has superior performance comparing with other machine learning algorithms.

The rest of this paper is organized as follows. Section 2 describes the dataset and the verification of its reliability. Section 3 presents the traffic accident severity prediction method based on the Deep Forests algorithm in this work, including the data preprocessing, which is of great importance for eliminating redundancies in the data and reorganizing the data efficiently. And the basic theory of feature selection and Deep Forests algorithm are introduced in this section as well. The experimental results are presented and discussed in Section 4, the application of this method is presented in Section 5, and conclusion and some future scopes are given in Section 6.

2. Data Description and Its Reliability Verification

This section firstly presents the data source adopted in this study. As this dataset has never been applied to severity prediction of a traffic accident, the reliability verification of this dataset is also conducted in this section.

2.1. Data Description. The analyses in this study are based on the road safety dataset of the United Kingdom in 2016. The data was obtained from the Kaggle website, a data prediction competition platform that allows data analysts to compete with each other to solve real and complex data science problems. The local characteristics of traffic accident data include 18 items in total, for example, longitude and latitude of the accident point, time characters of accident, type of the vehicle, gender of the driver, age of the driver, age of the vehicle, speed limit, light conditions, weather conditions, road surface conditions, and the other data characteristics. We use simple statistical analysis to perform a simple descriptive statistical analysis of the entire dataset. The age of driver ranges from 1 to 97 with an average of 36; the vehicle age is on average 5 ranging from 1 to 84 years. 70% of the drivers are male and others are female. The most vehicle type is car, accounting for 71%, followed by pedal cycle, occupying about 7%. As for the accident severity, about 85% are slight accident; fatal accidents account for only about 1%. Figure 1 shows the structure of this dataset.

2.2. Data Reliability Verification. As this dataset has never been applied to severity prediction of a traffic accident, the reliability verification of this dataset should be conducted before preprocessing of the data. Reasonable data distribution is an important manifestation of reliable data. Therefore, three dimensions (latitude and longitude distribution, date, and time) of data distribution are considered in this paper to verify the data reliability.

According to the latitude and longitude information of the original dataset, we use the visual plotting tools for intuitive analysis. Figure 2 shows the latitude and longitude distribution of the data, in which Figure 2(a) is a scatter plot based only on the longitude and latitude information of the dataset, while Figure 2(b) is obtained by matching the scatter plot with the real-world map. Through the visualization of data, we can obtain a general macroscopic understanding of the distribution of the entire accident data. Furthermore, we can easily find that the latitude and longitude information of the traffic accident is consistent with the map information, and there is no deviation beyond the range of the map, which indicates that the dataset is reliable in accident position distribution dimension.

Besides the location dimension, the “date” dimension and the “time” dimension are also two important dimensions for analyzing the dataset reliability. As for the measure index, we choose the month for the “date” dimension and week for the “time” dimension in this study. As is shown in Figure 3(a), the data is equally distributed through all months. From Figure 3(b), it is not difficult to find that the accidents occurred mostly on Friday, and the accidents on Saturday and Sunday were relatively mild, which is fully compatible with the actual situation. Additionally, in order to explore the law of traffic accident occurrence at a different time of the day, we separate the day’s hours from the “time” dimension and combine with the week index. The heat map of the accident occurring in different hours of one day is shown in Figure 4, from which we can find that most of the accidents occurred in the morning and evening peak hours of the working day. This is completely consistent with people’s travel characteristics during the weekday, which indicates that the data is therefore reliable.

3. Methodology

This section discusses the method used for our prediction study. To ensure and improve the prediction accuracy, data preprocessing including data cleansing and data normalization is carried out before the feature selection and severity prediction. Random Forests algorithm is applied to extract the significant features of traffic accidents based on the preprocessed data. Finally, the Deep Forests algorithm is applied to predict the severity of a traffic accident. The flow diagram of traffic accident severity prediction in this paper is depicted in Figure 5.

3.1. Data Correlation Verification. Before we use machine learning to predict the severity of an accident, we must confirm the necessity to choose the machine learning method to deal with such a problem. If the data is highly correlated, we can directly use the simpler linear model to directly predict, and then there is no need to use machine learning to solve the problem. Thus, we conduct the data correlation relationship verification in this section.

As well as giving details of date, time, and location, the dataset gives a summary of all reported vehicles and pedestrians involved in road accidents and other related accident features. 18 variables are taken into account in this paper, including accident severity, month of year, hour of day, vehicle reference, vehicle type, vehicle manoeuvre, journey purpose of driver, sex of driver, age band of driver, engine capacity, propulsion code, age of vehicle, driver home area type, day of week, speed limit, light conditions, weather conditions, and road surface conditions. The correlation relationship between all the features in the data is analyzed. As a consequence, a Pearson correlation matrix was plotted to identify the amount of linear relationship between variables and to determine whether linear-based algorithms are suitable through gaining insight into data. The matrix is color-coded, the numerical value one expressed in dark blue represents a completely positive linear correlation between two features, while turquoise represents a zero, suggesting no linear correlation. As is shown in Figure 6, the accident severity is independent of any of the other 17 features, which means that we cannot directly predict the accident severity with a simple linear model. Therefore, this paper considers a smarter machine learning approach to deal with this problem.

Additionally, it is worth noting that, in Figure 6, most of the characteristic variables are linearly independent, except for weather conditions, road surface, and light condition, the light condition and hour of day, vehicle type, and engine capacity. It can be easily and reasonably explained for these results. When it rained, the road conditions will become wet and the light condition will change to some extent. Similarly, with the advent of the night, light and environment will change according to the characteristics of time. Besides, different types of vehicles have different engine capacities. Therefore, the interactive relationship between these variables also proves the reliability of this dataset on the other hand.

Information of driver

Age_of_ driver	Sex_of_ driver	Journey_purpose_of_driver
45	1 (male)	2 (commuting to/from work)
21	2 (female)	1 (journey as part of work)
36	1 (male)	3 (taking pupil to/from school)
15	2 (female)	4 (pupil riding to/from school)
……	……	……

Information of road and environment

Speed_ limit	Light_conditions	Road_surface_ conditions	Weather conditions
30	1 (daylight)	2 (wet/damp)	1 (fine without high winds)
40	4 (darkness-lights lit)	1 (dry)	2 (raining without high winds)
20	6 (darkness-no lighting)	3 (snow)	3 (snowingwithout high winds)
50	5 (darkness-lights unlit)	4 (frost/ice)	7 (fog or mist)
……	……	……	……

Information of vehicle

Vehicle_type	Age_of_ vehicle	Engine_ capacity	Vehicle_manoeuvre
8 (taxi)	1	1896	9 (turning right)
4 (motorcycle)	15	689	2 (parked)
11 (bus or coach)	6	5883	11 (changing lane)
9 (car)	10	1995	18 (going ahead other)
……	……	……	……

Accident information

Accident_index	Date	Time	Day of week	Accident_severity
201506E098757	2015-03-09	12:56	2	3 (slight)
201506F006668	2015-07-04	21:33	7	1 (fatal)
201506F003976	2015-07-22	8:40	4	2 (serious)
……	……	……	……	……

Figure 1: Structure of the road safety dataset of the United Kingdom.

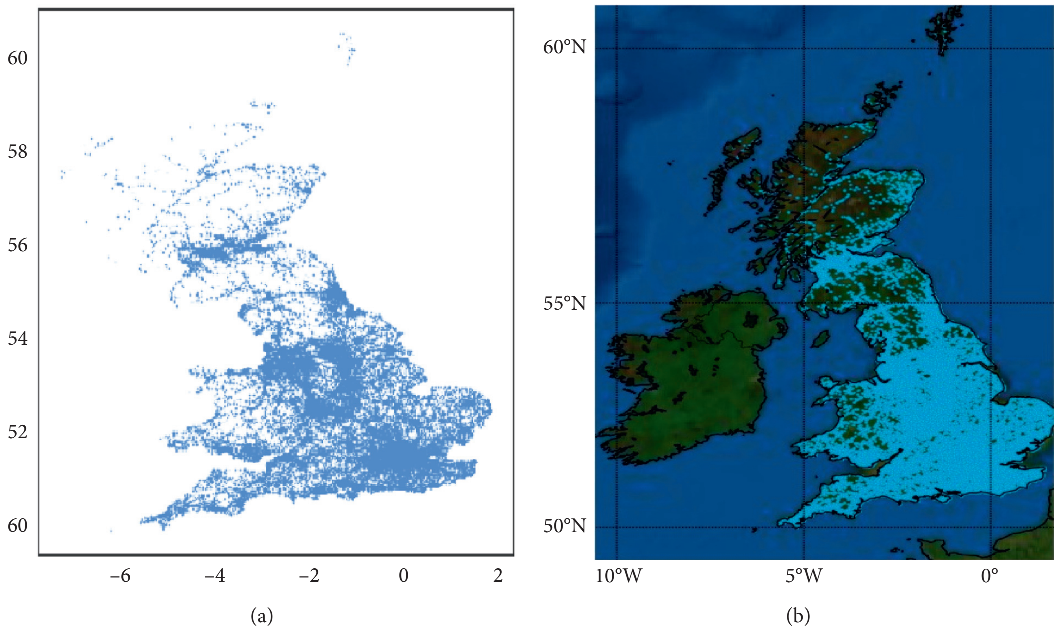

Figure 2: (a) The longitude and latitude map of the accident point and (b) the map-matching graph.

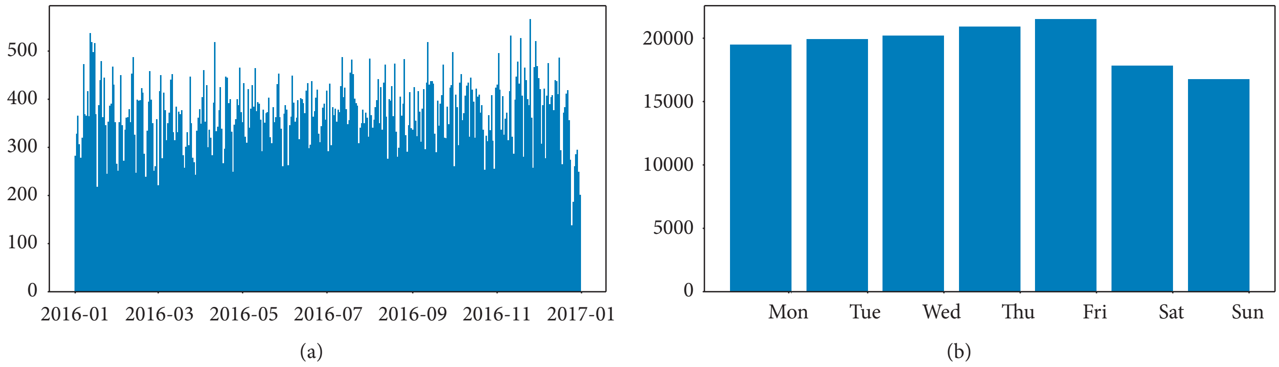

Figure 3: Date dimension and time dimension of the accident bar chart.

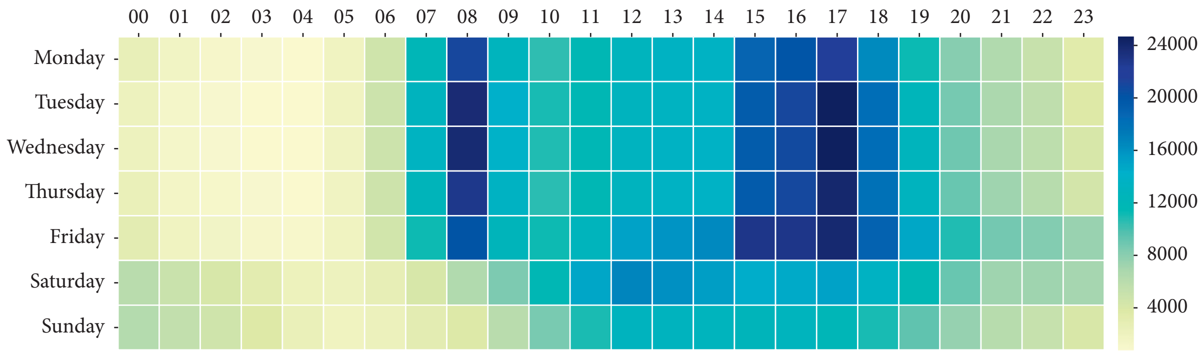

FIGURE 4: Traffic accident distribution per day and hour.

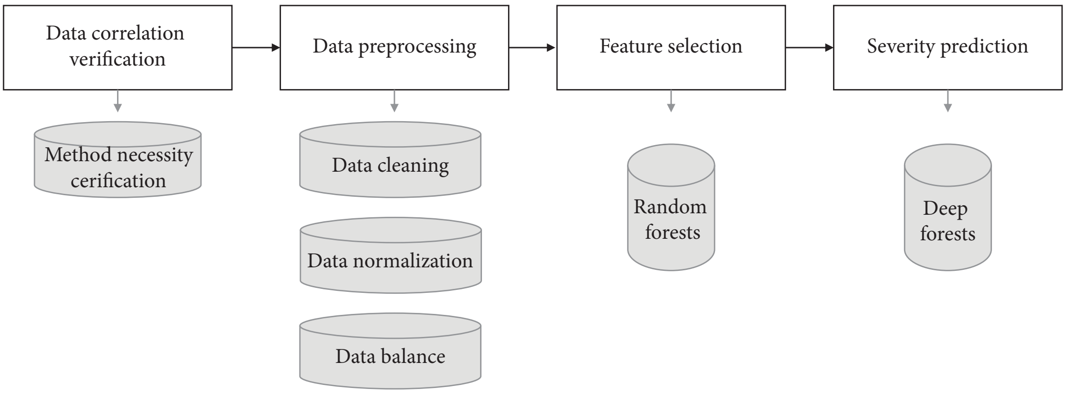

FIGURE 5: The flow diagram of traffic accident severity prediction method in this paper.

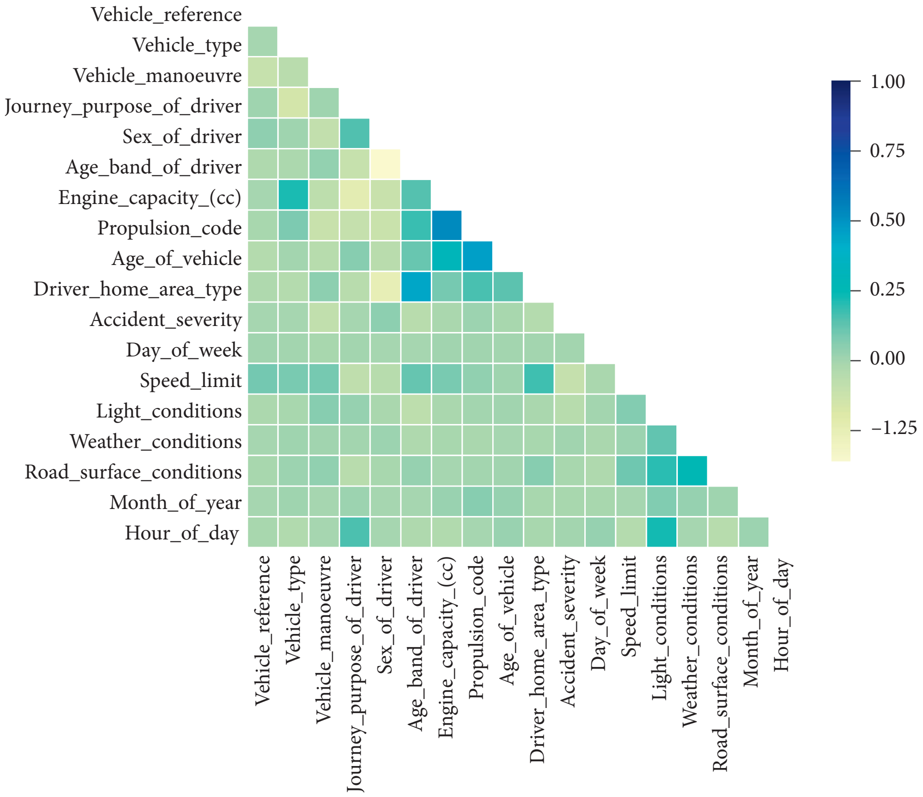

FIGURE 6: The Pearson correlation matrix of all features.

3.2. Preprocessing. It is of great importance to understand the nature of the available data and try to perform in-depth data analysis. Data preprocessing is very useful for meaningful data analysis; what we need to do is data cleaning, data normalization, and data selection in different class before our prediction analysis.

3.2.1. Data Cleaning. Data cleaning is the process of identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data from a record set, table, or database. The categorization criteria of all features in the dataset are listed, and the categorization criteria are defined by actual statistical results. So first, we need to observe the categorization criteria of each feature. However, due to the limited space, only a part of the categorization criteria of features is listed below. The categorization criteria of light conditions, weather conditions, and road surface conditions are shown in Table 1.

Through the statistical analysis of each feature of the original dataset, we found some obvious outliers and also some missing data that is labeled as "unknown" or "−1" needs to be cleaned up.

For some dimensions, the proportion of missing data exceeds 10%, and the average value replacement method was adopted. For example, in the dimension of Age_of_Vehicle, there are approximately 20% missing data labeled as "−1"; we adopted an average vehicle age of 5 to replace these missing values. For those dimensions with few missing data, we take a direct deletion method to clean them up, such as Road_Surface_Condition, where the missing data accounts for only 0.5%.

As for the obvious outliers, the same principle is adopted for the missing data processing method. For example, the Age_of_Driver is ranging from 1 to 97 with an average of 36; this age distribution is obviously unreasonable, because driving in the UK is only allowed for those over 17 years old. Because only 1% of the tags are under 17 years old, we thus directly delete them for the following processing.

3.2.2. Data Normalization. In the multi-index evaluation system, each evaluation index usually has different dimensions and orders of magnitude due to its different nature. When the levels between the indicators differ greatly if the analysis is performed directly with the original index values, the role of the higher-value indicators in the comprehensive analysis will be highlighted, and the effect of the low-level indicators will be relatively weakened. Therefore, in order to ensure the reliability of the results and to improve the convergence speed and accuracy of the model, the original indicator data needs to be normalized. Logarithm function conversion is adopted in this paper to conduct the normalization of all the given features to make sure features are on a similar scale. For example, for the feature Age of Vehicle, the age of the vehicle is between 1 and 84; the logarithmic method is used to standardize the distribution of the variable values so as to make the distribution of the variable values more "normal." Figure 7(a) depicts the distribution of Age_of_Vehicle before normalization, from which it can be easily found that the data shows obvious long tail characteristics. Normalization involves taking the logarithm of the given features. This is done because high values for certain variables computationally skew results more in favor of that variable than their actual contribution. In this case, age of the vehicle, for example, has values ranging from 1 to 84, when the majority of other categorical variables are binary or limited within 1–8 categories. After taking the log, one can notice that the values range from approximately 1 to 4, shown in Figure 7(b). This increases the performance of machine learning algorithms, as the numerical values do not have disproportionate amounts of computing value compared to all the other categorical variables.

3.2.3. Class Balance Verification. In the dataset, accident severity is listed as a classified label for prediction. Table 2 shows the criteria for categorizing accident severity and its distribution.

As can be seen from the distribution of data, the number of slight accidents is far greater than the number of fatal accidents, showing a long-tailed data distribution. In terms of model evaluation, accuracy was employed in this paper to compare the prediction performance. However, the accident severity level is unbalanced among three levels; therefore, the traditional classification algorithm with the overall classification accuracy as the learning goal will pay too much attention to the majority class, which will cause the accuracy paradox and deteriorate the classification performance of the minority class samples. This is why the data balance work should be conducted. The random sampling method was adopted in this paper. Both oversampling and undersampling have their own disadvantages, but this is the common problem of the imbalance of the dataset, which cannot be completely avoided.

After weighing the amount of data and enhancing the robustness of the model itself, we finally decided to take a combination of oversampling and undersampling to deal with this problem. Oversampling was adopted for training set to ensure as much training data as possible, trying repeated sampling to generate new rare samples to alleviate data imbalance. In addition, undersampling was adopted for test set to ensure that there are no duplicate samples in the test set, thereby improving the validity of the results.

After all this work was completed, 120,000 pieces of data for each category were obtained as the whole dataset. With the consideration of limited computational resources, 40000 pieces of data for each category were randomly selected as the training set and 2000 pieces of data for each category were screened out from the dataset as the test data for evaluating the performance of the model.

3.3. Feature Selection. An object usually has multiple properties, including related features, irrelevant features, and redundant features. Only these related features will improve the effectiveness of our learning algorithm. Since we are not aware which feature is effective for our prediction, dimensional disasters often occur in algorithmic

Table 1: Categorization criteria of several features.

Light conditions	Description	Weather conditions	Description	Road surface conditions	Description
1	Daylight: street lights present	1	Fine without high winds	1	Dry
2	Daylight: no street lighting	2	Raining without high winds	2	Wet/damp
3	Daylight: street lighting unknown	3	Snowing without high winds	3	Snow
4	Darkness: street lights present and lit	4	Fine with high winds	4	Frost/ice
5	Darkness: street lights present but unlit	5	Raining with high winds	5	Unknown
6	Darkness: no street lighting	6	Snowing with high winds		
7	Darkness: street lighting unknown	7	Fog or mist		
		8	Other		
		9	Unknown		

Table 2: Categorization criteria for traffic accident severity.

Accident severity code	Label	Distribution of the data
1	Fatal	2899 (1.12%)
2	Serious	34205 (13.27%)
3	Slight	220741 (85.61%)

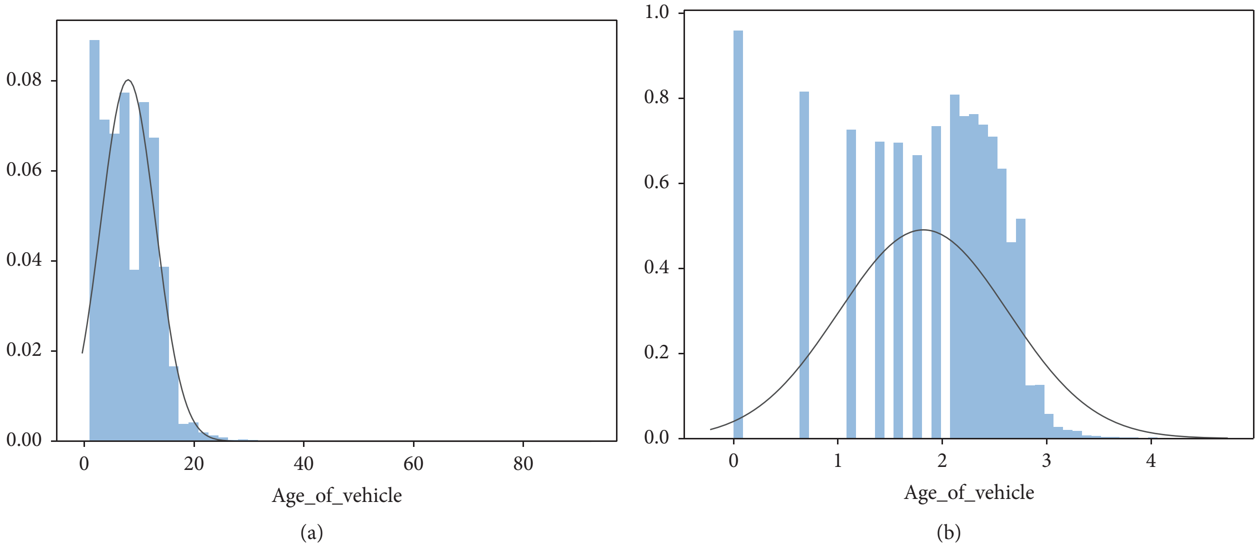

Figure 7: Data distribution (a) before normalization and (b) after normalization.

applications. So, it is of great significance to select relevant features from all features to improve the efficiency of the learning algorithm, especially for the analysis of complex data. A vast number of feature selection strategies have been proposed for applications in different fields [28–31]. In this paper, Random Forests (RFs) method is adopted to carry out feature selection according to the importance index of each feature, not only because of its ability to calculate the importance of a single feature variable, but also due to its good performance on most datasets.

RFs model is developed from decision-making regression trees, which will often generate hundreds of trees. The data of each tree is extracted from the bag of set B by bootstrap sampling method, while the remaining out-of-bag (OOB) samples are defined as set $\overline{B}$, which will not appear in the training samples. Let C define a set of B and $\overline{C}$ as a set of $\overline{B}$. Assuming $X_n \times p$ matrix is an n-dimensional test dataset with p characteristics, y is an n-dimensional label vector, and each value represents the corresponding category to which the test belongs. The random forest algorithm calculates the importance of the features by rearranging the errors before and after classification. Each feature X_j in the algorithm corresponds to a set of feature replacement tests with rearranged values. The importance of features is measured by comparing the classification error rates of the original features and the replaced randomly rearranged features in the OOB test set, which is the extent to which the change of original feature affects the result. When the important features are replaced by the randomly rearranged features, their discrimination will decrease; that is, the OOB classification error rate will increase. When N trees are established, there are N OOB sets as test sets. Therefore, the characteristic importance index J_a is defined as follows:

$$J_a\left(x_j\right) = \frac{1}{N} \sum_{\overline{B}_k \in \overline{C}} \frac{1}{\left|\overline{B}_k\right|} \left(\sum_{i \in B_k} I\left(h_k^{\overline{x}_j}(i) \neq y_i\right) - I\left(h_k(i) \neq y_i\right) \right), \tag{1}$$

where y_i is a classification label in the $i-th$ OOB, I denotes a characteristic function, $h_k(i)$ represents a classification label of sample i predicted by dataset B_k, and $h_k^{\overline{x}_j}(i)$ is a classification label after replacing characteristic x_j.

3.4. Severity Prediction. The representation learning in deep neural networks mainly depends on the processing of the original features by layer. Inspired by this, Zhou and Feng [18] obtained the cascade structure of Deep Forests as illustrated by the left schematic diagram in Figure 8. In a traditional deep neural network, each node denotes a neuron. In their research, the RFs were treated as a “forest neuron” and were stacked into multiple layers in deep learning. The cascade structure of deep neural networks is also presented by the right schematic diagram in Figure 8. Comparing with deep neural networks, the design concept of using Deep Forests resembles deep neural networks, and the “concatenate” and “vote” in Deep Forests resemble the nonlinear transformation procedures in deep learning. More significantly, the Deep Forests algorithm has much fewer hyper-parameters, each grade can be regarded as an ensemble of ensembles, and excellent performance is achieved in various domains by using the same parameter setting.

Each level of cascade receives feature information processed by its preceding level and outputs its processing result to the next level. Each level is an ensemble of decision trees forests, which means it can be regarded as an ensemble of ensembles. When a sample is given, each forest is calculated by calculating the percentage of different classes of training samples at the leaf nodes falling into the related instances, and then the average value of all the trees in the forest to generate the estimation of the distribution of the class. As shown in Figure 9, the red part highlights the path of each sample traversing leaf nodes. Different markings in leaf nodes represent different classes.

In order to reduce the risk of overfitting, the class vectors generated by each forest are generated by k-fold cross-validation. In particular, each instance will be used as the K-1 training data, producing a K-1 class vector, and then taking the average value to produce the final class vector as the enhancement feature at the lower level in the cascade. It is important to note that after a new level is extended, the performance of the entire cascade will be estimated on the validation set, and the training process will be terminated without significant performance gain. Therefore, the number of cascading cascades is automatically determined. Contrary to most deep neural networks with fixed complexity of the model, Deep Forests can determine the complexity of its model (early stop) properly through termination training, which enables Deep Forests to be applied to training data of different scales, not limited to large-scale training data.

4. Experimental Work and Results

This section introduces our experimental work and results with the methodology proposed in Section 3. To verify the superiority of our proposed method, several other machine learning algorithm-based perdition models were implemented to predict traffic accident severity with the same dataset, and the prediction results show that the Deep Forests algorithm with fewer hyper-parameters presents good stability and the highest accuracy under different level of training data volume.

4.1. Feature Selection. As described in 3.1, our dataset includes 18 features, and these features are almost independent of each other, which means that the complexity of this dataset is relatively high, and not all features are useful for improving forecasting accuracy since there may be some irrelevant or redundant features in those features. Therefore, before using the Deep Forests algorithm to predict the dataset, the feature selection work first is of great importance.

A combination of the Randomized Search and Grid Search method was adopted in this paper for parameter optimization. The Randomized Search method is applied firstly to quickly help us determine the approximate range of a parameter, and then we use the Grid Search method to cross-validate the selected candidate parameters of the model iteration and determine the optimal value of a parameter. The output of the best parameters is 5 for Max_depth, 2 for Min_samples_leaf, 10 for Min_samples_split, and 1000 for n_estimators. Therefore, a total of 1000 trees were used to grow the forest, and this number was deemed sufficient to yield reliable results. The feature importance ranking from the RFs is shown in Figure 10. Using the node purity measure, the explored variables were ranked in rising order from the least to the most important. Our principle for choosing the importance threshold is the ∅80 value of the cumulative value curve of importance. According to the importance value of each features, the ∅80 value is around 0.04; we thus adopted 0.04 as the critical value for the important features. Finally, eight features were chosen to conduct the accident severity prediction, including engine capacity, hour of day, age of vehicle, month of year, day of week, age band of driver, vehicle manoeuvre, and speed limit.

4.2. Severity Prediction Results. In this section, the eight features selected by the feature selection are used as the main data features. And then the Deep Forests algorithm is adopted to predict the severity of traffic accidents and produce the predicted accuracy. In our experiment, the cascade structure used in Deep Forests is as follows: each level consists of 4 completely random tree forests and 4

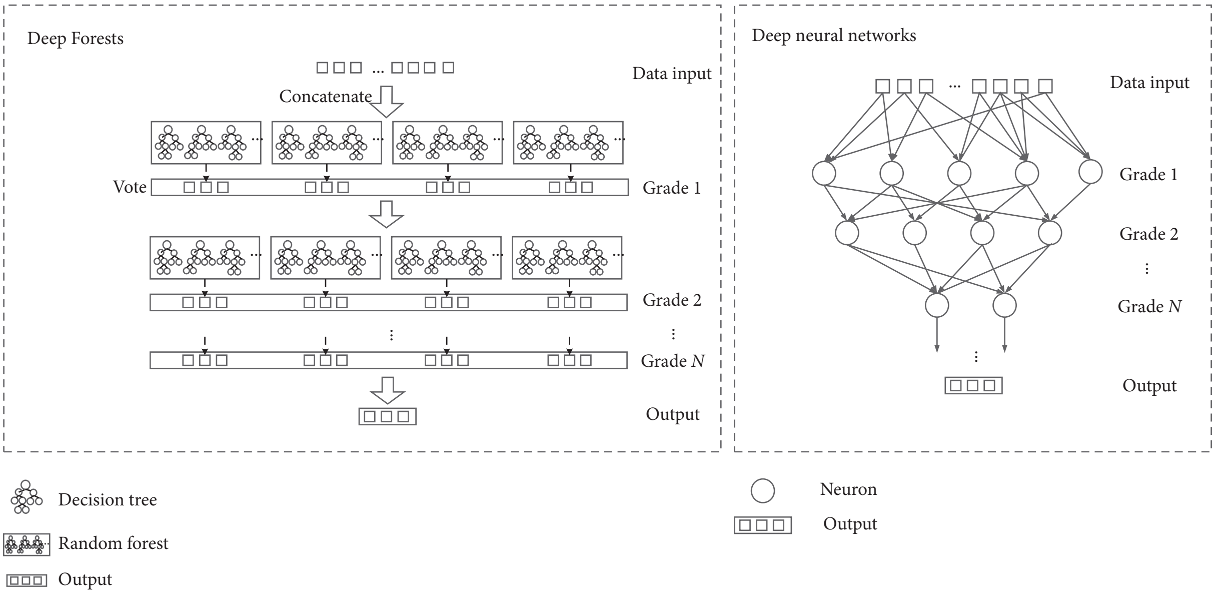

FIGURE 8: The cascade structure of Deep Forests and deep neural networks.

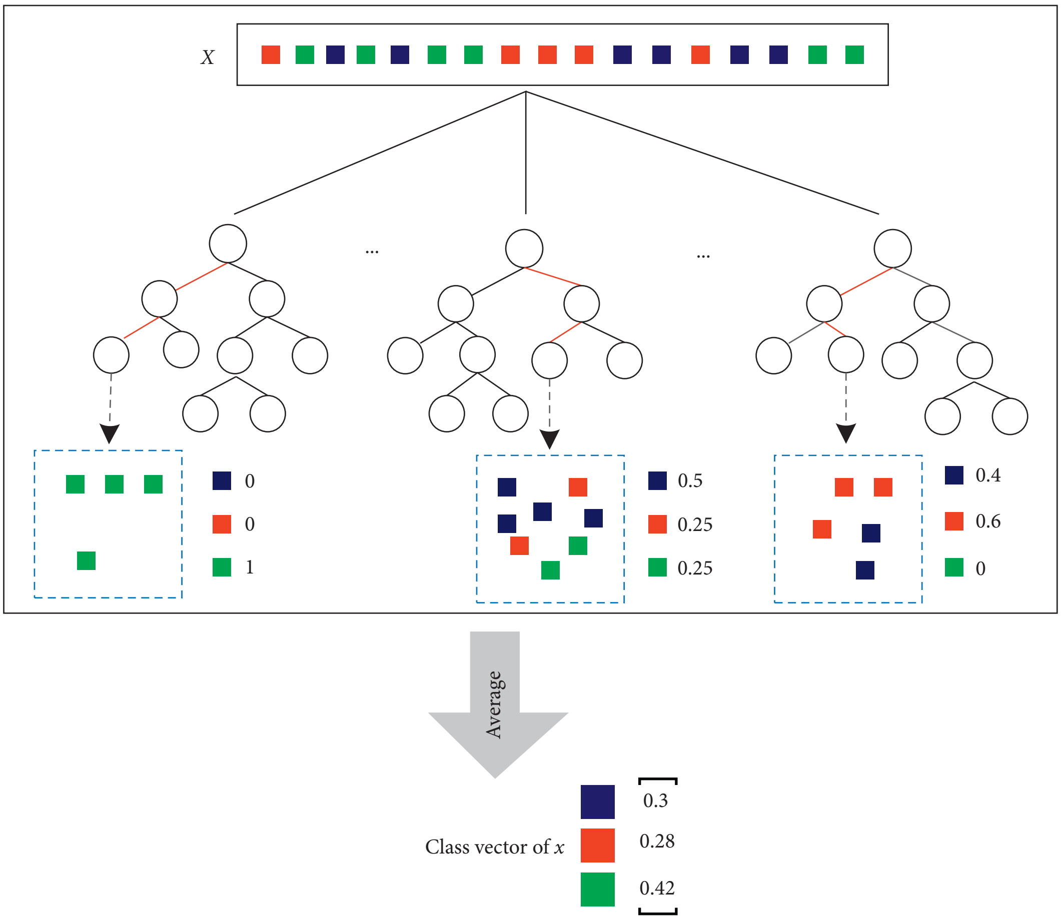

FIGURE 9: The classification process of training samples.

random forests, each with 500 trees, and three-fold CV is used for class vector generation. These settings of cascade structure are consistent with that proposed by Zhou and Feng [18], because it has been proven that this cascade structure is able to achieve excellent performance by using the same default setting in their paper. Hence, it is supposed

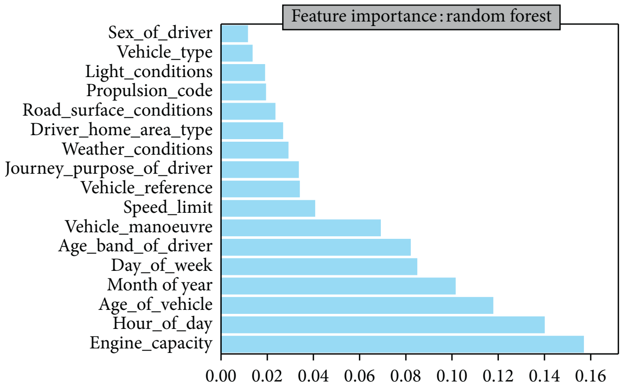

Figure 10: Feature importance results by Random Forests algorithm.

that this cascade structure is good enough with some consideration of performance and time consumption.

In order to verify that Deep Forests can achieve significant performance gains for traffic accident severity prediction, we compare Deep Forests with DNN and several other popular machine learning algorithms which are widely used in traffic accident prediction algorithms, such as Random Forests, LightGBM, XGboost, k-Nearest Neighbor (KNN), and decision trees. The computation progress for each algorithm is calculated and recorded by the same computer, which is equipped with a 2.8 GHz Intel Core i7 CPU and a 16 GB RAM. All of the forecasting models are implemented in Python language.

Table 3 illustrates the performance of Deep Forests, DNN, RFs, LightGBM, XGboost, KNN, and decision trees algorithms for traffic accident severity prediction. From the evaluation index results, Deep Forests algorithm performs better than other models. Recall is higher than other models; false alarm rate is lower than other models, so the overall F1 score is also higher. It shows that the model controls well the influence of data imbalance and learns the characteristics of different types of data. The ROC reached 90%, indicating that the model has learned the difference between different categories of data, and the prediction results are more reliable and stable.

In addition, the experimental results show that the direct use of DNN cannot achieve the desired effect on the problems studied in this paper. This is expected, because there are significant differences in the number of samples in different categories; it is difficult for the DNN model to learn the differences between categories. Without adding new data, we believe that constructing a more suitable deep learning model structure with careful tuned hyper-parameters can achieve better results to a certain extent, but this is beyond the scope of this paper. This is also the reason why this paper chooses the Deep Forests algorithm based on the characteristics of the dataset and the problem itself.

Due to the classification tasks of many data imbalance problems, we tend to pay more attention to the performance of the model on the minority class, the predictive performance of different accident categories is presented, as shown in Table 4. It can be easily found that the model performs worse in categories with fewer samples, compared with the predictive performance for the majority category. But the decline is less compared to other models, so the model adopted in this paper is more robust overall. In addition, in the performance of this imbalanced dataset, the tree-based models perform better than the neural network model; this is also the reason why we adopt Deep Forests model instead of the neural network model.

Table 3: Average predictive performance of different models.

	Accuracy (%)	Recall	False alarm rates	F1 score	Roc
Deep forests	90.69	0.92	0.09	0.91	0.93
RFs	88.98	0.90	0.10	0.90	0.92
XGboost	83.49	0.83	0.16	0.83	0.87
LightGBM	83.01	0.83	0.17	0.83	0.87
Decision tree	81.04	0.81	0.19	0.81	0.85
KNN	77.26	0.77	0.23	0.77	0.82
DNN	53.52	0.54	0.47	0.47	0.52

In order to better observe the performance of Deep Forests under different training data volumes, we divide the data into multiple orders of magnitude, and the accuracy of different magnitudes with different models are plotted in Figure 11, from which we can see that, with the increase of the sample size of the training set, the performance of each model has improved to a certain extent. However, the Deep Forests model is significantly better than other models at a small sample size, which also proves that the advantage of the model when dealing with small-scale sample size. In addition, the advantage of Deep Forests model is gradually weakened with the increase of sample size. When the sample size reaches 100,000, we can find that although the performance of Deep Forests is a little better than the random forest, it is not much different.

Additionally, compared with many traditional machine learning methods, the Deep Forests algorithm used in this paper has its own advantages. Deep Forests model has much fewer hyper-parameters than deep neural networks, although their iterative structure is similar. We usually do not know the optimal value of the model hyper-parameter for a given problem. Researchers generally rely on experience or use replicated values on other issues or search for the best values through trial and error. The increase in hyper-parameters will bring additional randomness to the model performance, which is too dependent on the regulation of hyper-parameters. For instance, there are many hyper-parameters in random forests that need to be constantly adjusted to optimize model prediction accuracy and speed up model calculations, including number of decision trees in the forest, the maximum number of features a random forest can have in a single tree, number of leaves, OOB sampling, and random state. However, the hyper-parameters in Deep Forests algorithm is less than random forests, and a set of hyper-parameters can be applied to different datasets as mentioned in literature [18], which is another big point of the deep forest algorithm used in this paper.

Table 4: The predictive performance of different accident categories.

	Recall			False alarm rates			F1 score			ROC		
	1	2	3	1	2	3	1	2	3	1	2	3
Deep Forests	0.93	0.82	1.00	0.17	0.09	0.01	0.88	0.86	1.00	0.91	0.88	1.00
RFs	0.91	0.77	1.00	0.19	0.10	0.01	0.86	0.83	1.00	0.90	0.86	1.00
XGboost	0.83	0.66	1.00	0.27	0.20	0.02	0.78	0.72	0.99	0.84	0.79	0.99
LightGBM	0.84	0.63	1.00	0.29	0.20	0.02	0.77	0.71	0.99	0.84	0.78	1.00
Decision tree	0.68	0.76	1.00	0.23	0.28	0.05	0.77	0.72	0.95	0.78	0.80	0.98
KNN	0.66	0.64	1.00	0.32	0.33	0.07	0.67	0.66	0.97	0.75	0.75	0.98
DNN	0.80	0.07	0.70	0.52	0.51	0.38	0.60	0.13	0.66	0.56	0.43	0.56

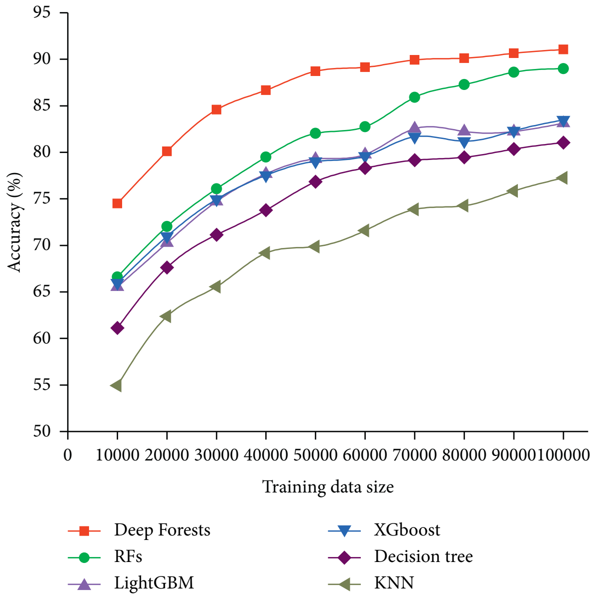

Figure 11: Accuracy rate comparison under different training data size.

5. Discussion

The higher prediction accuracy of our proposed method reveals that it can be used as a very useful tool for accident severity prediction. Fewer hyper-parameters in the deep forest will be more conducive to the transplantation of models; that is, a set of hyper-parameters can be applied to different datasets. Thus, it can be easily adapted to solve lots of different traffic problems as well, for instance, short-term forecast of travel time on expressway sections and traffic flow situation estimation. This is of great significance for the perfect improvement of the current traffic safety system within a sustainable transportation system, such as an intelligent transportation decision system and intelligent traffic safety management system.

From the perspective of traffic safety management implications, the more accurate severity prediction of traffic accidents has long been the research direction we are pursuing for sustainable transportation development. In most cases, many traffic safety control measures are still dominated by the limited experience of traffic managers, which may lead to a deviation from the actual situation. On the contrary, the use of many excellent deep learning algorithms can learn from the historical accident data record effectively and efficiently. The application of Deep Forests algorithms proposed in this paper has been proved to have good performance in predicting the severity of an accident. The prediction results can be used as an important and effective reference for the subjective judgment of safety managers. For instance, if a traffic safety management want to identify the important influencing factors of traffic accident and the severity level of traffic accidents caused by these factors, the general method we proposed in this paper can be easily carried out for different dataset by these managers to achieve their goals. In addition, the prediction outcomes of severity level can also provide an effective reference for the implementation of traffic accident management and control measures, such as the improvement of transportation infrastructure, the improvement of lighting conditions, the implementation of road variable speed limit, and driving safety warning.

6. Conclusions

With the recognition of the importance of machine learning in solving some problems in the transportation field, in this paper we innovatively apply the Deep Forests algorithm to the prediction of traffic accident severity. The excellent forecasting performance of our proposed method reveals that it can be used as a very useful tool for accident severity prediction. Fewer hyper-parameters in the deep forest will be more conducive to the transplantation of models; that is, a set of hyper-parameters can be applied to different datasets. Thus, it can be easily adapted to solve lots of different traffic problems as well, for instance, short-term forecast of travel time on expressway sections and traffic flow situation estimation, although from the analysis results there is still room for improvement in prediction accuracy. This is because we have not done enough in the mining of raw data. For future research of this study, in order to improve prediction accuracy, we will try to summarize and construct some features that do not exist in the features of raw data based on the information of the data features. In addition, it should be noted that this paper does not focus on optimizing the model parameters, which is also a research direction in the future. Nevertheless, the method proposed in this paper has certain contributions to both theory and practice.

Data Availability

The data used in this article is from the open-source database Kaggle, which can be downloaded freely at https://www.kaggle.com/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Key R&D Program of China (Grant no. 2018YFC0704704).

References

[1] World Health Organization (WHO), *A Road Safety Technical Package*, World Health Organization, Geneva, Switzerland, 2017.

[2] X. Meng, L. Zheng, and G. Qin, "Traffic accidents prediction and prominent influencing factors analysis based on fuzzy logic," *Journal of Transportation Systems Engineering and Information Technology*, vol. 2, p. 15, 2009.

[3] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1666–1676, 2011.

[4] K. Li, D. Qian, S. Huang, and X. Liang, "Analysis of traffic accidents on highways using latent class clustering," in *Proceedings of the 19th COTA International Conference of Transportation Professionals-CICTP*, pp. 1800–1810, Nanjing, China, July 2016.

[5] B. Dadashova, B. A. Ramírez, J. M. McWilliams, and F. A. Izquierdo, "The identification of patterns of interurban road accident frequency and severity using road geometry and traffic indicators," *Transportation Research Procedia*, vol. 14, pp. 4122–4129, 2016.

[6] Z. Yan, X. Lu, and W. Hu, "Analysis of factors affecting traffic accident severity based on heteroskedasticity ordinal Logit," in *Proceedings of the Sixth International Conference on Transportation EngineeringICTE 2019,* American Society of Civil Engineers Reston, pp. 422–435, Chengdu, China, September 2020.

[7] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Analysis & Prevention*, vol. 88, pp. 37–51, 2016.

[8] B. García de Soto, A. Bumbacher, M. Deublein, and B. T. Adey, "Predicting road traffic accidents using artificial neural network models," *Infrastructure Asset Management*, vol. 5, no. 4, pp. 132–144, 2018.

[9] X.-F. Zhang and L. Fan, "A decision tree approach for traffic accident analysis of saskatchewan highways," in *Proceedings of the 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Toronto, Canada, May 2013.

[10] Z. Pu, Z. Li, Y. Jiang, and Y. Wang, "Full bayesian before-after analysis of safety effects of variable speed limit system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–13, 2020.

[11] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," *IEEE Access*, vol. 6, pp. 60079–60087, 2018.

[12] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," in *Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, pp. 361–364, Hangzhou, China, December 2017.

[13] P. Mudali, J. Roopa, M. G. Raju, and A. Yadav, "Analysis of Parallel M5P and Random Forest Regression for Visualization of Traffic Behavior," in *Proceedings of the Computational Intelligence in Pattern Recognition*, pp. 231–241, Springer, Hangzhou, China, December 2020.

[14] P. Nadarajan, M. Botsch, and S. Sardina, "Predicted-occupancy grids for vehicle safety applications based on autoencoders and the random forest algorithm," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1244–1251, IEEE, Anchorage, AK, USA, January 2017.

[15] J. Li, "Bus arrival time prediction based on random forest," in *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*, Taiyuan, China, June 2017.

[16] O. H. Kwon and S. H. Park, "Identification of influential weather factors on traffic safety using K-means clustering and random forest," *Advanced Multimedia and Ubiquitous Engineering*, vol. 393, pp. 593–599, 2016.

[17] Y. R. W. Gao Z and A. Ke, "Urban expressway traffic incident duration prediction based on random survival forests," *Journal of Tongji University*, vol. 45, no. 9, pp. 1304–1310, 2017.

[18] Z. Zhou and J. Feng, "Deep forest: towards an alternative to deep neural networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, August 2017.

[19] S. Rota Bulo and P. Kontschieder, "Neural decision forests for semantic image labelling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 81–88, Columbus, OH, USA, June 2014.

[20] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1467–1475, Santiago, Chile, December 2015.

[21] G. Biau, E. Scornet, and J. Welbl, "Neural random forests," *Sankhya A*, vol. 81, no. 2, pp. 347–386, 2019.

[22] Y. Wang, S.-T. Xia, Q. Tang, J. Wu, and X. Zhu, "A novel consistent random forest framework: bernoulli random forests," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3510–3523, 2017.

[23] M. S. Horswill and S. Helman, "A behavioral comparison between motorcyclists and a matched group of non-motorcycling car drivers: factors influencing accident risk," *Accident Analysis & Prevention*, vol. 35, no. 4, pp. 589–597, 2003.

[24] E. K. Adanu, S. Jones, and K. Odero, "Identification of factors associated with road crashes among functionally classified transport modes in namibia," *Scientific African*, vol. 7, Article ID e00312, 2020.

[25] Z. Pu, S. Wang, C. Liu, Z. Cui, and Y. Wang, "Road surface friction prediction using long short-term memory neural network based on historical data," *Electrical Engineering and Systems Science*, https://arxiv.org/abs/1911.02372, 2020.

[26] M. Touahmia, "Identification of risk factors influencing road traffic accidents," *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2417–2421, 2018.

[27] Z. Pu, Z. Cui, S. Wang, Q. Li, and Y. Wang, "Time-aware gated recurrent unit networks for forecasting road surface friction

using historical data with missing values," *IET Intelligent Transport Systems*, vol. 14, no. 4, pp. 213–219, 2020.

[28] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[29] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[30] J. Ou, J. Xia, Y.-J. Wu, and W. Rao, "Short-term traffic flow forecasting for urban roads using data-driven feature selection strategy and bias-corrected random forests," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2645, no. 1, pp. 157–167, 2017.

[31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.