IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 22, 2020, accepted November 4, 2020, date of publication November 16, 2020,
date of current version November 25, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3037922*

# Application of XGBoost for Hazardous Material Road Transport Accident Severity Analysis

**XIAOYAN SHEN[1,2] AND SHANSHAN WEI[1]**
[1]School of Automobile, Chang'an University, Xi'an 710064, China
[2]Key Laboratory for Automotive Transportation Safety Enhancement Technology of the Ministry of Communication, Chang'an University, Xi'an, China

Corresponding author: Shanshan Wei (weishan1995@foxmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFE0108000.

**ABSTRACT** Hazardous material road transport accidents pose a serious threat to public life, property and the environment. Therefore, studying the factors influencing road transport accidents involving hazardous materials can help identify the main causes behind them and contribute to the adoption of specific and targeted measures to reduce casualty rates and improve traffic safety. However, most existing research either adopted methods based on statistical analysis or neglected to further evaluate the spatial relationships. This study aims to use the eXtreme Gradient Boosting (XGBoost) algorithm to analyze hazardous material road transport accident data from seven regions of China. Considering the rarity of these events, the classification performance of different methods is compared based on precision, recall, F-score and Area Under Curve (AUC). The results indicate that the proposed XGBoost method has the best modeling performance. There is some variation between regions in the features that have a significant impact on accident severity. The influence of the same feature on the severity of an accident even varies from region to region. The aforementioned results provide a theoretical basis for exploring the issues, sustainability, challenges, and tasks of safe transportation activities for hazardous materials in the future. These results can help regions develop targeted prevention and response policies to efficiently reduce the incidence and severity of accidents.

**INDEX TERMS** Feature analysis, hazardous materials, road transport accident, transport safety, XGBoost.

## I. INTRODUCTION

In recent years, with the continuous development of China's economy, the market demand for hazardous chemicals has increased, of which 95% of hazardous materials come from different places than their destination and more than 50% are transported by road [1]. In 2018, the volume of hazardous chemicals transported by road in China reached 1.86 billion tons. The rapid increase in the frequency of transportation has led to a significant rise in the frequency of hazardous chemical road transport accidents. In addition, since hazardous materials are flammable, explosive, corrosive and poisonous, accidents often lead to more serious secondary injuries, causing a series of social problems, such as damage to the ecological environment and increased casualties and property losses. According to the U.S. Department of Transportation's report on hazmat accidents (2009-2018), 145,971 (87.90%) of the 166,065 hazmat accidents occurred while in transit, and the number of highway-related incidents is increasing every year (12,730 in 2009 and 17,923 in 2018) [2]. In China, from 2006 to 2017, 3,974 incidents involving the transport of hazardous materials resulted in the loss of 5,203 lives. This finding indicated that more than one person was killed every day in China as a result of a hazardous materials incident [3]. On March 1, 2014, two tanker trucks carrying methanol collided in a road tunnel in Shanxi Province, causing a fire that killed 40 people. On June 13, 2020, a vehicle transporting liquefied petroleum gas exploded during transport in Zhejiang Province, causing the collapse of nearby houses and factories, killing 20 people and injuring 175 others [1].

Over the past decades, the issue of hazardous material transportation has been a very active area of research. However, most studies have focused on the direct costs of hazardous material transport or quantifying the potential losses that can result from an accident [4]–[6]. Research on the factors influencing the severity of hazardous material road transport accidents has been limited. Moreover, most research has merely described the characteristics of an accident or explored the relationship between the features and the severity based on statistical methods. Statistical models can quantitatively describe the functional relationship between a phenomenon and certain factors. Andersson, an early pioneer in the study of hazmat accidents, used statistical methods to analyze 570 hazmat accidents from

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

1986 to 1987 and determined that the type of hazardous material, type of road, type of truck for transportation, and location of the area had a significant impact on the severity of the accidents [7]. Yang *et al*., 2010 conducted a statistical survey of hazmat road transport accidents that occurred in China from 2000-2008 and found that 46.6% of the accidents were caused by poor road conditions, 13.7% were caused by driver error and 9% were caused by mismanagement [8]. Zhang *et al*. found that 1632 accidents involving hazmat trucks occurred in China from 2006 to 2010; the majority of them resulted in hazardous material spills, followed by explosion (15.1%) and fire (5.3%), and leakage was often the cause of subsequent explosion or fire [9]. A total of 708 accidents involving hazardous materials on Chinese highways from 2004 to 2011 were analyzed by Shen *et al*. [10]. Their study identified that 56% of those accidents resulted in hazardous material spills and that vehicle defects and human error were the main causes of hazmat accidents. Ma *et al*. [11] used an ordered logit model to estimate the probability of hazmat accidents of different severity levels and applied elasticity theory to analyze the factors significantly influencing the severity of hazmat accidents. They found that the factors that dramatically influence the severity of road hazmat accidents are illegal behavior, unsafe driving behavior, accident responsibility, vehicle problems, vehicle type, weather, lighting, road level, and regional distribution. Duan [12] analyzed hazardous chemical accidents in China for the period spanning 2000 to 2006 and found that the more developed southeast coastal areas had a higher incidence of accidents and deaths than other regions. Poku-Boansi *et al*. [13] found that vehicle speed, the presence of a spill and the population density at the accident road had a significant effect on the severity of road transport accidents involving dangerous goods. A random parameters ordered probit model was established by Xing *et al.* [14] to explore the influence of contributing factors on the severity of accidents. The results indicated that higher injury severity may be related to hazmat type, mishandling, driver fatigue, speeding, tunnels, slopes, county roads, dry roads, winter, darkness, more than two vehicles, rear-end accidents, and explosions. The results of a study by Fabiano *et al.* [15] showed that road alignment, meteorological factors and the frequency of transport vehicle traffic significantly affect the risk of road transport of dangerous goods. Azimi *et al.* [16] employed a random parameter logit model to study the injury severity of large truck rollover crashes in the state of Florida, and they identified that crashes tend to be more severe when there are hazardous material spills.

Statistical models have been used to successfully explore the factors influencing the severity of traffic accidents. A statistical model is an a priori hypothesis about the potential relationship between the variables of interest to determine the effect of the independent variable on the dependent variable after understanding the statistical characteristics, such as the data collection method and the estimated quantity. However, in practice, there is a possibility that the a priori assumptions do not represent the real situation of the variables, leading to inappropriate inferences [17]. In addition, related studies have pointed out that statistical models are more suitable for exploring the relationships embodied in data with smaller sample sizes and narrower characteristics [18], [19].

In contrast, machine learning models, as nonparametric tools, do not assume relationships between endogenous and exogenous variables and have no or few presuppositions about the explanatory variables. These models are more adaptable and can process high-dimensional data quickly; the larger the sample size is, the better the analysis. Furthermore, these models have the ability to classify dependent variables by calculating the highest significant explanatory variables [20], [21]. Currently, machine learning research is focused on decision trees (DTs), random forests (RFs), artificial neural networks (ANNs), support vector machine (SVM), etc. [22]–[24]. However, algorithmic processes such as ANNs and SVM are performed as if in a black box, making it difficult to see the process and directly obtain differences in the effect of different features on accident severity [25]. Tree-based algorithms are a common approach in machine learning algorithms. These algorithms have progressed from single decision trees to random forests based on bagging algorithms to gradient boosting trees. In a continuous improvement process, eXtreme Gradient Boosting (XGBoost) has improved the basic framework of a gradient boosting machine (GBM) by optimizing the system and enhancing the algorithm to offset all parallelization overheads in computation [26]. Additionally, borrowing regular terms corrects the inherent overfitting of a tree model. Ultimately, XGBoost has demonstrated the distinctive capability to solve a variety of classification problems and is widely recognized among researchers for its accuracy, simplicity, and interpretability. Soleimani *et al.* [27] used XGBoost to determine the relative importance of the variables used to close a crossing based on accident data occurring at 18,485 road-rail grade crossings in the United States. The model accuracy was 0.991, which was higher than that of decision trees (0.984) and random forests (0.987). Bahador *et al.* [28] applied XGBoost and SHapley Additive exPlanations (SHAP) for real-time accident detection and characterization. The results showed that XGBoost can robustly detect accidents with 99% accuracy, 79% detection and a 0.16% false alarm rate. It was also proposed that characteristics such as speed, population, network, land use, and weather conditions had a significant impact on the probability of accidents. Ma *et al.* [26] conducted a spatial analysis of the leading factors for the 3,146 traffic fatalities that occurred in Los Angeles in 2010-2012 based on a methodological framework of XGBoost and grid analysis and identified eight factors as the most influential. The influences were, in descending order, drunk driving, involvement in parties, rear-end collisions, lighting conditions, pedestrian involvement, motorcycle involvement, day of the week, and time of day. Zhang *et al.* [29] modeled the hierarchical relationship between material properties and their deep semantics occurring in the same image by the GS-XGBoost

algorithm, which has been applied in different scenarios such as large-scale product image retrieval, robotics, and industrial inspection. Shi *et al.* [30] applied XGBoost to urban fire incident prediction.

In general, the literature on the analysis of factors influencing the severity of hazardous material road transport accidents is limited and has mainly focused on accident descriptions using statistical methods, with few studies applying machine learning algorithms to the analysis of factors influencing the severity of hazardous material road transport accidents. In addition, the previously small sample size and failure to account for inter-regional variability has created knowledge gaps in identifying key influences and predicting crash severity. The purpose of this paper is to analyze the factors influencing the severity of hazardous material road traffic accidents in seven regions of China. The severity of an accident was divided into property damage only, injury and fatal, depending on the casualty. The nonparametric machine learning algorithm XGBoost was applied in this paper for data preprocessing and exploration of key risk features, and its performance was compared with that of four other common classification algorithms. The comparison showed that XGBoost outperforms the other algorithms in terms of classification accuracy. The knowledge gained from this study can provide a theoretical basis for the government and transport enterprises to formulate effective preventive measures, rescue programs and material reserve plans to minimize a series of social problems, such as casualties, property damage and environmental pollution.

The remainder of this paper is organized as follows:

Section II provides an introduction to the XGBoost algorithm. Section III describes the data sources and processing procedures. Section IV presents the results of the model assessment and data analysis, and improvement recommendations are made based on the results of the data analysis.

## II. METHODOLOGY

### A. XGBoost

XGBoost is a C++ optimized implementation of a GBM [26], [31], [32]; complexity is introduced into the model when measuring the efficiency of the algorithm, so the objective function of XGBoost is expressed as:

$$Obj = \sum_{i=1}^{m} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{1}$$

where $i$ represents the $i$th sample in the dataset, $m$ indicates the total amount of data imported into the $k$th tree, and $K$ stands for all trees created. When only $t$ trees are created, the equation should be $\sum_{k=1}^{t} \Omega\left(f_k\right)$. $y_i$ represents the true label, $\hat{y}_i$ represents the predicted value, and $\Omega$ represents an equation that measures the complexity of the tree model from the structure of the tree.

When $t$ trees are created, the predicted value $\hat{y}_i$ in the traditional loss function can be expressed in the following manner:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t-1} f_k\left(x_i\right) + f_t\left(x_i\right) = \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \tag{2}$$

It follows that the traditional loss function is related to all trees that are well established. $\hat{y}_i$ contains the results of all tree iterations [29], thereby establishing a direct link between the structure of the tree and the model effect. The objective function can be expressed as:

$$Obj = \sum_{i=1}^{m} l\left(y_i^{(t)}, \hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \sum_{k=1}^{t-1} \Omega\left(f_k\right) + f_t \tag{3}$$

The objective function can be expressed as follows after expansion based on Taylor's formula:

$$Obj = \sum_{i=1}^{m} \left[ l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right) + f_t\left(x_i\right) g_i + \frac{1}{2}\left(f_t\left(x_i\right)\right)^2 h_i \right] + \sum_{k=1}^{t-1} \Omega\left(f_k\right) + \Omega\left(f_t\right) \tag{4}$$

where $g_i = \frac{\partial l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}}$ and $h_i = \frac{\partial^2 l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)}{\partial^2 (\hat{y}_i^{(t-1)})}$ are the first- and second-order derivatives of the loss function $l(y_i^{(t)}, \hat{y}_i^{(t-1)})$ over $\hat{y}_i^{(t-1)}$, respectively.

The constant term is irrelevant to the result of the $t$th iteration, so the constant terms $l\left(y_i^{(t)}, \hat{y}_i^{(t-1)}\right)$ and $\sum_{k=1}^{t-1} \Omega\left(f_k\right)$ are removed from the objective function. The objective function can be expressed as:

$$Obj = \sum_{i=1}^{m} \left[ f_t\left(x_i\right) g_i + \frac{1}{2}\left(f_t\left(x_i\right)\right)^2 h_i \right] + \Omega\left(f_t\right) \tag{5}$$

The structure of the tree is redefined according to formula (6),

$$f_t\left(x_i\right) = w_{q(x_i)} \tag{6}$$

where $q\left(x_i\right)$ denotes the leaf node where sample $x_i$ is located. $w_{q(x_i)}$ denotes the score obtained by this sample falling in the $q\left(x_i\right)$ leaf node of the $t$th tree.

If a tree contains a total of $T$ leaf nodes, where the index of each leaf node is defined as $j$, then the weight of the samples on the leaf nodes is $w_j$. The complexity of the model $\Omega\left(f\right)$ can be expressed as:

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

By bringing the structure of the tree into the loss function and defining the set of samples contained on a leaf with index $j$ as $I_j$, the objective function can be transformed into the following equation (8):

$$Obj = \sum_{j=1}^{T} \left[ w_j \sum_{i \in I_j} g_i + \frac{1}{2} w_j^2 \left( \sum_{i \in I_j} h_i + \lambda \right) \right] + \gamma T \tag{8}$$

### B. CROSS-VALIDATION

In $k$-fold cross-validation, the training set is split into $k$ subsets. For each of the $k$ "folds", the following procedure is followed. A model is trained using the $k-1$ folds as training data. The resulting model is validated on the remainder of the data (i.e., these data are used as a test set to compute performance measures, such as accuracy). The $k$-fold cross-validation then reports a performance measure that is the average of the values computed in the loop.

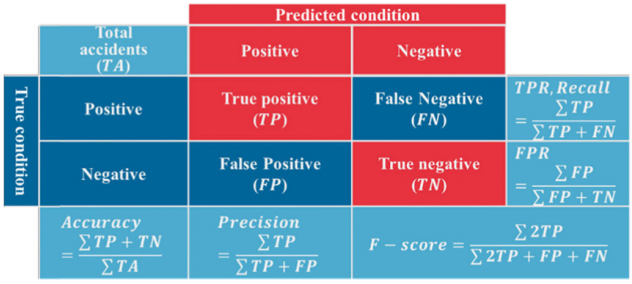


**FIGURE 1.** Confusion matrix and formulas for calculating accuracy, TPR, FPR, precision, recall, and F-score.

This method is computationally costly but does not waste much data, which is a tremendous advantage in problems with very small sample sizes. Previous tests have shown that the use of cross-validation improves the results of the model [33], and 10-fold cross-validation is widely used.

### C. MODEL ASSESSMENT INDICATORS

#### 1) CONFUSION MATRIX

The confusion matrix and the metrics associated with it, accuracy, true positive rate (TPR), false positive rate (FPR), precision, recall, F-score, receiver operating characteristic (ROC), and the area under the ROC curve (AUC), were used to evaluate the model in this study [34].

A confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class, making it easy to see the numbers of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). This approach allows more detailed analysis than the mere proportion of correct guesses (accuracy). The four outcomes and calculation formulas for assessing indicators are shown in Figure 1 as follows:

#### 2) ACCURACY

Accuracy is a composite metric that reflects how many of all samples are correctly predicted and is one of the most commonly used metrics for assessing predictive performance in classification mandates. In general, the higher the accuracy rate is, the better the classifier.

#### 3) TRUE POSITIVE RATE (TPR) AND FALSE POSITIVE RATE (FPR)

TPR indicates the proportion of samples that the classifier predicts to be positive as a percentage of the number of samples that are actually positive and measures the classifier's ability to identify positive examples. FPR expresses the proportion of samples that the classifier predicts to be positive among the actual number of negative samples.

#### 4) PRECISION, RECALL AND F-SCORE

Precision can be defined as a measurement of accuracy, i.e., the proportion of positive samples that are predicted to be correct among the total number of samples predicted to be positive.

Recall is a metric of completeness, i.e., the number of positive samples predicted correctly as a percentage of the number of actual positive samples. F-score is the harmonic mean of precision and sensitivity. The best values for precision, recall and F-score are close to 1, and the worst values are close to 0 [35].

#### 5) RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE AND AREA UNDER THE ROC CURVE (AUC)

ROC is a curve with FPR as the horizontal coordinate and TPR as the vertical coordinate, and this curve reflects a combination of the continuous variables of sensitivity and specificity. The larger the AUC is, the better the diagnostic performance [36].

## III. DATA

### A. DATA COLLECTION

The data used in this paper were collected by the Dangerous Chemicals Registration Center of the Ministry of Emergency Management of the People's Republic of China. The data represent the occurrence of road transport accidents involving hazardous materials in seven regions of China over the five-year period from 2015-2019. Based on the real situation of the raw data and with reference to the factors affecting the safety of road transport of dangerous goods listed in the European Agreement concerning the International Carriage of Dangerous Goods by Road (EUR), Highway Routing of Hazardous Materials: Guidelines for Applying Criteria (U.S.), and Regulations on the Administration of Dangerous Chemicals Safety (CN) documents, 19 features were initially selected as the independent variables of the model. These features are accident forms (direct accident form: DAF, final accident form: FAF), driver attributes (qualification: QU, fatigue: FAT), vehicle attributes (vehicle type: VT, vehicle safety status: VSS, device security status: DSS, moving state: MS), road attributes (road type: RT, road alignment: RA, traffic signal: TS, intersection: INT, segment type: ST), environmental attributes (surface condition: SC, season: SEA, month: MON, time of day: TOD, weather: WEA), and type of hazardous materials: HM. The severity level of an accident was determined by the number of casualties and was divided into three levels (property damage only: O, injury: I, and fatal: F).

### B. DATA PREPROCESSING

The complexity of hazardous material road transport incidents and the lack of specialization in the collection of information on hazardous material road transport incidents mean that there are always shortcomings in our raw data, and preprocessing is often required before the data can be applied to the model. The preprocessing process in this study involved data cleaning and data formatting.

#### 1) DATA CLEANING

Highly relevant data will be removed from the dataset [37]. Highly relevant data (correlation coefficient above 0.5) include data that are strongly correlated with the target and data that are tightly correlated with each other.

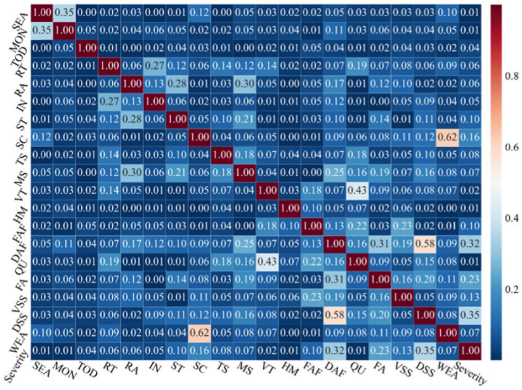


**FIGURE 2.** Description of multicollinearity between variables.

The results of a correlation analysis of the data are displayed in Figure 2, where one of each pair of features with a correlation coefficient greater than 0.5 is removed to alleviate the correlation problem and reduce the computational cost. In summary, this study excluded two highly correlated features. The number of features was reduced to 17. The cleaned dataset and its attribute descriptions are listed in Table 1. The number of hazardous material road transport accidents in each region of China is shown in Figure 3.

#### 2) DATA FORMATTING

Most of the features collected in this study are not sequential but categorical nominal variables, for which only the use of dummy variables can convey the most accurate information possible to the algorithm [38]. In the data, season, road type, etc., are nominal variables that need to be converted to dummy variables using unique hot coding. For example, in the raw dataset, the categorical data for SEA have four independent labels, including spring, summer, fall and winter. After a one-hot encoder was applied, four dummy variables, SEA_1, SEA_2, SEA_3, and SEA_4, were given to indicate the season in which an accident occurred. In this way, the 17 categorical features were formatted into 91 dummy object variables.

## IV. RESULTS AND DISCUSSION

### A. XGBOOST MODEL

The experiment was run on a computer with 8 GB of running memory, an Intel (R) Core (TM) i3-3110M CPU, and a Windows 10 operating system. The coding environment was Python 3.8.2.

#### 1) MODEL PERFORMANCE ASSESSMENT

To further test the performance of XGBoost, four popular models, logistic regression (LR), multilayer perceptron (MLP), random parameters logit model (RPLM), random forest (RF) and SVM, were used to compare the performance, and 10-fold cross-validation was used to stabilize the results. The results are shown in Figure 4.

#### 2) XGBOOST PERFORMANCE ANALYSIS

The results describing the performance of the classifier for the seven regions, calculated from the confusion matrix, are shown in Table 2.

East China, Northwest China and Central China are the regions with more hazardous material road transport accidents in China, and the performance of the model for those regions is superior to that for the other regions. This finding may be due to the fact that there are fewer accident records in the other regions. These results clearly demonstrate that the model may not obtain the desired predictive accuracy when the dataset is too small.

### B. FEATURE IMPORTANCE

The combination of feature importance and XGBoost's decision rules allows for a more definitive and comprehensive exploration of the main features that have an impact on the severity of hazardous material road transport accidents in each region. Specific effective measures and suggestions can be proposed to enhance the safety of hazardous material road transport. The main features affecting the severity of hazardous material road transport accidents in different regions are listed in Figure 5. Table 3 lists the occurrences of accidents (property damage only, injury, and fatal) with and without the relevant features. More specific details will be discussed in the next section.

### C. FEATURE ANALYSIS

The impact of each characteristic on the severity of hazardous material road transport accidents in the local area is analyzed based on the main risk characteristics of each region.

#### 1) EAST CHINA

The following results can be obtained from Figure 5 and Table 3. In East China, the features that have the greatest influence on the severity of hazardous material road transport accidents include HM, SC, MS, FA, and TOD (in order of importance).

Road transport accidents involving Class III and VIII hazardous materials accounted for 78% of all accidents, and the probabilities of serious and major accidents were higher than those of other types of hazardous materials (I: 48% VS 38%; F: 8% VS 7%). Frequent transport may be the underlying cause, and the flammable, explosive and corrosive properties of hazardous substances increase the likelihood of a serious accident [9].

Accidents occurring on dry pavement accounted for 83% of the total accidents. The casualty probability of accidents that occurred on dry pavement is significantly lower than that of the other road surface conditions (I: 41% VS 66%; F: 6% VS 14%). This is mainly because the high friction coefficient of dry pavement enables drivers to prevent an accident in time. However, on wet pavement with a lower coefficient of friction, the adhesion between a vehicle and the pavement is less, and it is not easy to control vehicles, which exacerbates the seriousness of an accident [14]. East China is located in China's eastern coastal and southeastern regions, with a temperate and subtropical monsoon climate, more rain in summer, and more snow in winter, further exacerbating the above situation.

Sixty-one percent of the total number of accidents occurred while the vehicle was traveling straight ahead.

**TABLE 1.** Descriptive statistics of features.

| Feature | Code and Description | Count | % | Feature | Code and Description | Count | % |
|---|---|---|---|---|---|---|---|
| SEA | 1: Spring | 324 | 23.0 | ST | 1: Ordinary segment | 1111 | 78.7 |
| | 2: Summer | 428 | 30.3 | | 2: Bridge | 56 | 4.0 |
| | 3: Autumn | 367 | 26.0 | | 3: Tunnel | 45 | 3.2 |
| | 4: Winter | 292 | 20.7 | | 4: Entrance and exit | 54 | 3.8 |
| MON | 1: January | 86 | 6.1 | | 5: Station | 88 | 6.2 |
| | 2: February | 68 | 4.8 | | 6: Risky segment | 40 | 2.8 |
| | 3: March | 112 | 7.9 | | 7: Other | 17 | 1.2 |
| | 4: April | 105 | 7.4 | MS | 1: Go straight | 733 | 51.9 |
| | 5: May | 106 | 7.5 | | 2: Turn | 467 | 33.1 |
| | 6: June | 97 | 6.9 | | 3: Shunting | 103 | 7.3 |
| | 7: July | 164 | 11.6 | | 4: Downhill | 16 | 1.1 |
| | 8: August | 166 | 11.8 | | 5: Stop | 92 | 6.5 |
| | 9: September | 129 | 9.1 | DAF | 1: Leakage | 216 | 15.3 |
| | 10: October | 126 | 8.9 | | 2: Fire | 85 | 6.0 |
| | 11: November | 114 | 8.1 | | 3: Explosion | 7 | 0.5 |
| | 12: December | 138 | 9.8 | | 4: Rollover | 401 | 28.4 |
| TOD | 1: [1-2] | 144 | 10.2 | | 5: Running out of road | 11 | 0.8 |
| | 2: [3-4] | 100 | 7.1 | | 6: Colliding with a fixed object | 128 | 9.1 |
| | 3: [5-6] | 79 | 5.6 | | 7: Fall down | 20 | 1.4 |
| | 4: [7-8] | 162 | 11.5 | | 8: Two-vehicle rear collision | 313 | 22.2 |
| | 5: [9-10] | 222 | 15.7 | | 9: Two-vehicle collision | 185 | 13.1 |
| | 6: [11-12] | 182 | 12.9 | | 10: Multivehicle rear collision | 17 | 1.2 |
| | 7: [13-14] | 54 | 3.8 | | 11: Multivehicle collision | 15 | 1.1 |
| | 8: [15-16] | 202 | 14.3 | | 12: Other | 13 | 0.9 |
| | 9: [17-18] | 83 | 5.9 | FAF | 1: Leakage | 1214 | 86.0 |
| | 10: [19-20] | 19 | 1.3 | | 2: Fire | 146 | 10.3 |
| | 11: [21-22] | 82 | 5.8 | | 3: Explosion | 42 | 3.0 |
| | 12: [23-24] | 82 | 5.8 | | 4: Rollover | 9 | 0.6 |
| SC | 1: Dry | 1181 | 83.7 | VT | 1: Tank lorry | 1247 | 88.4 |
| | 2: Wet | 136 | 9.6 | | 2: Cargo truck | 140 | 9.9 |
| | 3: Waterlogged | 46 | 3.3 | | 3: Other | 24 | 1.7 |
| | 4: Ice | 48 | 3.4 | VSS | 1: Safety | 1280 | 90.7 |
| TS | 1: YES | 1368 | 97.0 | | 0: Malfunction | 13` | 0.9 |
| | 0: NO | 43 | 3.0 | QUA | 1: Yes | 1334 | 94.5 |
| RT | 1: Freeway | 648 | 45.9 | | 0: No | 77 | 5.5 |
| | 2: National highway | 226 | 16.0 | FAT | 1: Yes | 316 | 22.4 |
| | 3: Provincial road | 187 | 13.3 | | 0: No | 1095 | 77.6 |
| | 4: Rural road | 124 | 8.8 | HM | 1: Explosives | 1 | 0.1 |
| | 5: Urban road | 226 | 16.0 | | 2: Gases | 274 | 19.4 |
| RA | 1: Straight | 795 | 56.3 | | 3: Flammable liquid | 837 | 59.3 |
| | 2: Winding road | 437 | 31.0 | | 4: Flammable solids | 22 | 1.6 |
| | 3: Ramp | 109 | 7.7 | | 5: Oxidizers and organic peroxides | 25 | 1.8 |
| | 4: Long downhill | 11 | 0.8 | | 6: Poisonous and infectious substances | 25 | 1.8 |
| | 5: Other | 59 | 4.2 | | 8: Corrosives | 226 | 16.0 |
| INT | 1: Yes | 226 | 16.0 | | 9: Miscellaneous | 1 | 0.1 |
| | 0: No | 1185 | 84.0 | | | | |

The probability of a fatal crash occurring when the vehicle is traveling straight ahead is less than that of other moving states (F: 6% VS 10%). The primary reason is that when going straight, the driver is already relatively well acquainted with

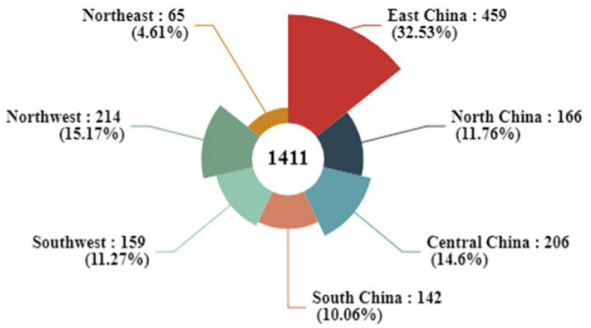


**FIGURE 3.** Accident distribution in different districts.

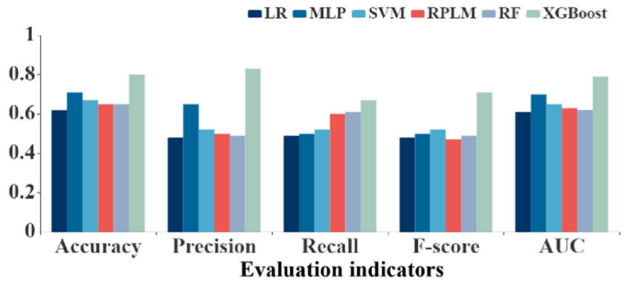


**FIGURE 4.** Comparison of models.

**TABLE 2.** XGBoost performance metrics for the 7 districts.

| District | Accuracy | Precision | Recall | F-score | AUC | Class |
|---|---|---|---|---|---|---|
| East China | 0.95 | 0.90 | 0.88 | 0.89 | 0.88 | O |
| | | 0.79 | 0.93 | 0.85 | 0.88 | I |
| | | 0.97 | 0.30 | 0.46 | 0.88 | F |
| East China | 0.79 | 0.77 | 0.77 | 0.77 | 0.83 | O |
| | | 0.81 | 0.85 | 0.83 | 0.83 | I |
| | | 0.00 | 0.00 | 0.00 | 0.83 | F |
| East China | 0.81 | 0.88 | 0.65 | 0.75 | 0.85 | O |
| | | 0.76 | 0.94 | 0.84 | 0.85 | I |
| | | 0.95 | 0.60 | 0.75 | 0.85 | F |
| East China | 0.74 | 0.70 | 0.94 | 0.80 | 0.79 | O |
| | | 0.80 | 0.71 | 0.75 | 0.79 | I |
| | | 0.90 | 0.20 | 0.30 | 0.79 | F |
| Southwest | 0.77 | 0.97 | 0.50 | 0.67 | 0.80 | O |
| | | 0.72 | 0.97 | 0.84 | 0.80 | I |
| | | 0.00 | 0.00 | 0.00 | 0.80 | F |
| Northwest | 0.83 | 0.89 | 0.67 | 0.76 | 0.86 | O |
| | | 0.80 | 0.97 | 0.88 | 0.86 | I |
| | | 0.98 | 0.50 | 0.67 | 0.86 | F |
| Northeast | 0.70 | 0.60 | 0.60 | 0.60 | 0.76 | O |
| | | 0.73 | 0.92 | 0.81 | 0.76 | I |
| | | 0.00 | 0.00 | 0.00 | 0.76 | F |

the surrounding environment and can deal with a potential accident that is happening in time. However, when turning, avoiding or going downhill, the road transport environment is relatively complicated and unfamiliar; these conditions not

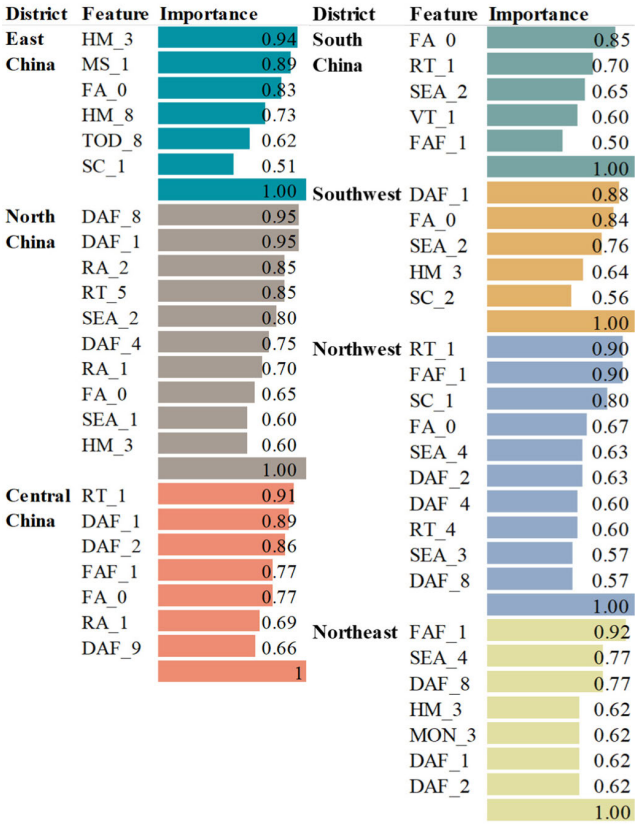


**FIGURE 5.** The significance of features.

only increase the accident rate but also increase the severity of an accident if the driver fails to respond in a timely manner [16].

Twenty-three percent of all accidents in the region occurred with drivers who were fatigued while driving. Injuries and fatalities are more likely to occur in a fatigued state than in a non-fatigued state (I: 64% VS 40%; F: 12% VS 6%). The reason for this is that when fatigued, drivers are slow to become aware and react. When an accident is imminent, a fatigued driver is unable to assess the danger or take the correct avoidance measures in time. This dramatically increases the number of potential fatalities in traffic accidents [39].

Accidents at 15:00 and 16:00 in the afternoon accounted for 15% of the total number of accidents, which was significantly higher than the average. The rate of fatal accidents was higher than that at other times of day (F: 9% VS 7%). The main reasons include the fact that East China exerts strict control over the transportation times of hazardous materials, and these measures reduce the accident rates at night and in the early morning. After driving for a long time, truck drivers become fatigued at 15:00-16:00 and lose the ability to judge the driving environment around them, thus increasing the severity of accidents [40].

### 2) NORTH CHINA

From Figure 5 and Table 3, we can reach the following findings. In North China, DAF, RT, FA, RA, SEA, and HM are

**TABLE 3.** Casualties for the selected features.

| District | Involve these features | | | | | Not involve these features | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| East China | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| HM (3, 8) | 44 | 48 | 8 | 358 | 78 | 55 | 38 | 7 | 101 | 22 |
| SC (1) | 53 | 41 | 6 | 379 | 83 | 20 | 66 | 14 | 80 | 17 |
| MS (1) | 48 | 46 | 6 | 281 | 61 | 44 | 46 | 10 | 178 | 39 |
| FA (0) | 54 | 40 | 6 | 353 | 77 | 24 | 64 | 12 | 106 | 23 |
| TOD (8) | 47 | 44 | 9 | 70 | 15 | 47 | 46 | 7 | 389 | 85 |
| North China | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| DAF (1,4,8) | 37 | 61 | 2 | 129 | 78 | 46 | 51 | 3 | 37 | 22 |
| RT (5) | 60 | 40 | 0 | 25 | 15 | 35 | 62 | 2 | 141 | 85 |
| FA (0) | 44 | 56 | 1 | 133 | 80 | 21 | 73 | 6 | 33 | 20 |
| RA (1,2) | 34 | 64 | 2 | 140 | 84 | 65 | 35 | 0 | 26 | 16 |
| SEA (1,2) | 43 | 56 | 1 | 86 | 52 | 35 | 63 | 3 | 80 | 48 |
| HM (3) | 38 | 60 | 3 | 104 | 63 | 42 | 58 | 0 | 62 | 37 |
| Central China | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| DAF (1,2,9) | 65 | 29 | 5 | 75 | 36 | 21 | 69 | 10 | 131 | 64 |
| RT (1) | 38 | 48 | 14 | 77 | 37 | 36 | 59 | 5 | 129 | 63 |
| FAF (1) | 33 | 61 | 6 | 176 | 85 | 60 | 17 | 23 | 30 | 15 |
| FA (0) | 44 | 49 | 7 | 151 | 73 | 18 | 71 | 11 | 55 | 27 |
| RA (1) | 41 | 48 | 11 | 115 | 56 | 32 | 64 | 4 | 91 | 44 |
| South China | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| FAF (1) | 45 | 50 | 5 | 121 | 85 | 57 | 19 | 24 | 21 | 15 |
| VT (1) | 44 | 49 | 7 | 123 | 87 | 68 | 21 | 11 | 19 | 13 |
| FA (0) | 55 | 42 | 4 | 108 | 76 | 24 | 56 | 21 | 34 | 24 |
| RT (1) | 40 | 50 | 10 | 86 | 61 | 59 | 38 | 4 | 56 | 39 |
| SEA (2) | 46 | 38 | 16 | 50 | 35 | 48 | 49 | 3 | 92 | 65 |
| Southwest | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| DAF (1) | 90 | 10 | 0 | 21 | 13 | 23 | 70 | 7 | 138 | 87 |
| FA (0) | 33 | 63 | 4 | 131 | 82 | 29 | 57 | 14 | 28 | 18 |
| SEA (2) | 35 | 56 | 9 | 57 | 36 | 30 | 66 | 4 | 102 | 64 |
| HM (3) | 31 | 65 | 4 | 102 | 64 | 33 | 58 | 9 | 57 | 36 |
| SC (2) | 21 | 58 | 21 | 19 | 12 | 34 | 63 | 4 | 140 | 88 |
| Northwest | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| DAF (2,4,8) | 28 | 64 | 8 | 118 | 55 | 49 | 36 | 15 | 96 | 45 |
| RT (1,4) | 38 | 48 | 14 | 120 | 56 | 37 | 55 | 7 | 94 | 44 |
| FAF (1) | 39 | 55 | 7 | 183 | 86 | 29 | 32 | 39 | 31 | 14 |
| SC (1) | 41 | 51 | 8 | 188 | 88 | 12 | 54 | 35 | 26 | 12 |
| FA (0) | 43 | 47 | 10 | 167 | 78 | 17 | 66 | 17 | 47 | 22 |
| SEA (3,4) | 35 | 50 | 15 | 112 | 52 | 40 | 53 | 7 | 102 | 48 |
| Northeast | O% | I% | F% | Count | % | O% | I% | F% | Count | % |
| FAF (1) | 8 | 77 | 15 | 47 | 72 | 44 | 50 | 6 | 18 | 28 |
| SEA (4) | 50 | 50 | 0 | 14 | 22 | 33 | 57 | 10 | 51 | 78 |
| DAF (1,2,8) | 50 | 47 | 3 | 32 | 49 | 24 | 64 | 12 | 33 | 51 |
| HM (3) | 33 | 55 | 13 | 40 | 62 | 44 | 56 | 0 | 25 | 38 |
| MON (3) | 14 | 57 | 29 | 7 | 11 | 40 | 55 | 5 | 58 | 89 |

identified as key determinants of severity in accidents, in that order.

In 78% of the accidents, the direct accident forms involved spills, rollovers and two-vehicle rear-end collisions. Accidents with the above direct accident forms were less likely to result in fatalities than those with other direct accident forms (F: 2% VS 3%). Possible causes include the following: in comparison to the direct forms of accidents described above, an explosion leaves very little time for people to escape. Fall down accidents generally occur

on treacherous roads or at bridges, making rescue difficult and thus increasing the severity of an accident. Multivehicle accidents involve a large number of people, which in turn increases the potential fatality rate. Similarly, in multivehicle accidents, those who are not initially injured and decide to flee their vehicle immediately are still at risk [10].

Fifteen percent of the total crashes occurred on urban roads. The probabilities of injury and fatal levels for hazardous material road transport accidents the occurred on urban roads were less than those of other road types (I: 40% VS 62%; F: 0% VS 2%), which is related to the strict regulation of the time of entry of vehicles transporting hazardous materials on urban roads. More serious accidents on highways and national and provincial roads can be attributed to the high speed of traffic, complicated traffic mix and number of parties involved. Lower police control and the poor road transport environment on rural roads increase the probability of fatal accidents [11].

Twenty percent of the total accidents occurred when drivers were fatigued. Accidents were more likely to be fatal when the driver was fatigued (F: 6% VS 1%). This finding may be due to the fact that fatigued drivers are slower to become aware and react [41].

The casualty probability for hazardous material road transport accidents that occurred on straight and curved roads was significantly higher than that for the other road alignments (I: 64% VS 35%; F: 2% VS 0%). Furthermore, 84% of accidents occurred on straight and curved roads. This result can be attributed to the fact that the main road alignments in North China are straight and curved roads. Driving on straight roads for long periods of time can cause visual fatigue, or driving on straight roads can be too comfortable and increase the likelihood of negligent driving, which can lead to serious accidents. At curves, a large mass of fluid in a tank can easily lead to overturning due to inertia when making turns, leading to casualties [15].

Fall and winter have higher fatality rates than spring and summer (F: 3% VS 1%). In North China, the need for heating during the fall and winter months leads to a significant rise in demand for hazardous materials and frequent transportation, which in turn increases the potential fatality rate [10]. Moreover, the cooler temperatures in autumn and winter pose a challenge to the technical safety of vehicles and equipment [3].

Accidents involving the transport of Class III hazardous materials have a higher probability of casualties than those of other types of hazardous materials (I: 60% VS 58%; F: 3% VS 0%). This finding may be attributed to the larger proportion of Class III hazardous materials (63%).

#### 3) CENTRAL CHINA

The following can be derived from Figure 5 and Table 3. In Central China, DAF, RT, FAF, FA, and RA, in that order, are critical features in determining the severity of road transport accidents related to hazardous materials.

When direct accident forms involve spills, fires, and two-vehicle collisions, the occurrence probability of injury or death is lower than that in other direct accident forms (I: 29% VS 69%; F: 5% VS 10%). The reasons for this result are similar to those in the previous section (North China). However, only 36% of the accidents in this region involved the above direct accident forms.

Accidents on highways accounted for 37% of all accidents in the region and were more likely to result in fatalities than accidents on other types of roads (F: 14% VS 5%). This finding might be attributed to more vehicles on the highway leading more easily to multivehicle accidents; the more parties involved in an accident, the higher the number of people to be engaged and the higher the fatality rate. Moreover, it is prevalent that the higher the speed is, the higher the mortality rate [42].

Eighty-five percent of accidents in the region ended in a spill as the final form of the accident. Accidents where the final accident form was a spill were associated with a much lower probability of fatalities than that of other final accident forms (F: 6% VS 23%). This may be explained by the fact that if the final accident form is a spill, the accident may mostly be caused by the failure of equipment and not involve other vehicles. In addition, leaks leave more escape time for accident participants than rollovers, fires, or explosions.

In 27% of the accidents in the region, drivers were fatigued at the time of the accident. Fatigued driving revealed higher rates of injury and death (I: 71% VS 49%; F: 11% VS 7%). This result is also caused by the poor condition of a driver when fatigued.

The influence of road alignment on accident severity is mainly reflected in the probability of fatal accidents. Fatalities are approximately 2.75 times more likely to occur on straight roads than on other road alignments (F: 11% VS 4%), mainly because drivers are more relaxed when driving on straight roads, making them more prone to drowsy or careless driving [43]. Worse still, accidents on such road alignments accounted for 56% of the region's accidents.

#### 4) SOUTH CHINA

From Figure 5 and Table 3, we can reach the following findings. In South China, the severity of hazardous material road transport accidents basically depends on FAF, VT, FA, RT, and SEA, in that order.

Accidents in which the final accident form was a spill were more likely to involve injuries and significantly less likely to result in fatalities than those with other final accident forms (I: 50% VS 19%; F: 5% VS 24%); the reasons are consistent with those in the previous section [9]. In addition, 85% of the accidents ended up in the form of a spill.

Tanker trucks accounted for 87% of all hazardous material road transport vehicles. Accidents involving tanker trucks had a significantly higher probability of injury accidents and a lower probability of fatal accidents than those involving other vehicle types (I: 49% VS 21%; F: 7% VS 11%). The reasons

behind this are as follows: tankers are the main vehicles used to transport hazardous materials, the regulation of tankers is becoming more systematic, the design and manufacture of tankers are more sophisticated, and the safety level of vehicles is increasing [44]. Other vehicles are mostly illegal transport vehicles that evade regulations; for these vehicles, the equipment safety level is not up to standard, and the driver has a lack of knowledge of hazardous chemical road transport and rescue, increasing the probability of fatal accidents [11].

Twenty-four percent of drivers were fatigued at the time of the accident. Fatigued drivers were more likely than non-fatigued drivers to be involved in both injury and fatal accidents (I: 56% VS 42%; F: 21% VS 4%). This finding is caused by the poor condition of a driver when fatigued.

Accidents on freeways accounted for 61% of all accidents in the region, and the probabilities of injury and fatal levels for crashes that occurred on freeways were greater than those of other road types (I: 50% VS 38%; F: 10% VS 4%). The possible explanations for the above results are as follows. First, because the highway road environment is better, when driving on such a road, a driver will unknowingly increase speed; second, more vehicles on the highway can easily lead to multivehicle accidents; the more parties that are involved in an accident, the higher the number of people to be engaged and the higher the fatality rate [16].

In summer, accidents were more likely to be fatal (F: 16% VS 3%), and 35% of accidents occurred in the summer. These findings are mainly due to the fact that summer is the main season for road transport of hazardous materials, which increases the possibility of accidents due to frequent transport. Additionally, high temperatures and heavy rainfall have a great impact on the transport environment, the technical safety of vehicles and equipment and the attention of drivers, further increasing the chances of serious accidents [45].

#### 5) SOUTHWEST CHINA

The following results can be obtained from Figure 5 and Table 3. In the Southwest, features that have a significant impact on the severity of road transport accidents involving hazardous materials include the DAF, FA, SEA, HM, and SC, in order of importance.

When the direct accident form was a spill, the severity of an accident was significantly lower than those of other direct accident forms, and the probability of a fatal accident was zero (I: 10% VS 70%; F: 0% VS 7%). This may be because spills are usually caused by the aging of equipment or by a minor impact, which will not readily lead to serious accidents. However, accidents where the direct accident form was a spill accounted for only 13% of the total number of accidents.

According to our results, fatigue has a substantial effect on the incidence of fatal accidents. Mortality in a fatigued state is 3.5 times higher than that in a non-fatigued state (F: 14% VS 4%), and the interpretation of this result is the same as presented previously. Eighteen percent of drivers were fatigued at the time of an accident.

The season in which an accident occurs has a dramatic impact on fatalities. Fatal accidents were more likely to occur in summer than in other seasons (F: 9% VS 4%), with 36% of accidents in the region occurring during the summer months. This is primarily attributed to the fact that summer is the season with the most frequent transportation of hazardous materials, the traffic volume is large, the number of parties involved in accidents is large and the number of people involved is large, thus increasing the probability of fatal accidents [18], [46]. On the other hand, summer precipitation is more frequent in Southwest China, with 78% of days experiencing precipitation and a large amount of precipitation, approximately 300 mm. Persistent heavy rain reduces road conditions and a driver's ability to observe the surrounding environment, increases the tension of driving and affects the driver's ability to control the vehicle, and the likelihood of a serious accident in this state is greater [45].

The probability of a fatal accident involving Class III hazardous materials was significantly smaller than that of other types of hazardous materials (F: 4% VS 9%). In the Southwest, 64% of accidents involved the transport of Class III hazardous materials. Unlike other regions, the Southwest had a lower probability of fatalities from transporting Class III hazardous materials than from transporting other types of hazardous materials, and this finding can be traced to the region's strict regulations on transporting Class III hazardous materials.

Road surface conditions have a high correlation with the probability of fatal accidents. The probability of fatal crashes occurring on wet pavement was greater than that on other road surface conditions (F: 21% VS 4%), and 12% of accidents in this region involved wet road conditions [45]. In addition to the abovementioned climatic reasons, the complex geographical environment of the Southwest region causes certain difficulties for rescue, which is also a reason for the high rate of fatal accidents.

#### 6) NORTHWEST CHINA

From Figure 5 and Table 3, we can draw the following conclusions. In Northwest China, the severity of hazardous material road transport accidents is mainly influenced by DAF, RT, FAF, SC, FA, and SEA (in order of importance).

Accidents involving the direct accident forms of fires, rollovers and two-vehicle rear-end collisions accounted for 55% of the total number of accidents in the region. The probability of a fatal accident was lower under the above direct accident forms than under other direct accident forms (F: 8% VS 15%). The reasons for this result are analogous to those in the North China region.

Accidents involving highways and rural roads accounted for 56% of the total number of accidents in the region. The probability of fatal hazardous material road transport accidents that occurred on highways and rural roads was significantly higher than that on other road types (F: 14% VS 7%). Possible reasons include high speeds, large numbers of vehicles and complex road conditions on highways, as well

as lax transport management and poor road infrastructure on rural roads, preventing timely rescue efforts [47].

Spills accounted for the largest proportion of final accident forms (86%) and were far less likely to be fatal than other final accident forms (F: 7% VS 39%). An explanation of this result can be found in the section on the Central China region.

Eighty-eight percent of accidents in this region occurred on dry road surface conditions. Accidents on dry roads were significantly less likely to be fatal than those on wet, waterlogged or icy roads (F: 8% VS 35%). The possible reasons for this result are as follows: the Northwest has more plateau and mountainous terrain with difficult terrain and poor road conditions, reducing driver control and the opportunity to adjust the vehicle when the road surface is wet or icy. Additionally, lower levels of emergency response and medical care increase the probability of fatal accidents [45].

Twenty-two percent of accidents in the region occurred when drivers were fatigued. Injuries and fatalities were more likely to occur in fatigued conditions than in non-fatigued conditions (I: 66% VS 47%; F: 17% VS 10%) for the same reasons as before.

Fifty-two percent of all accidents in the region occurred during the fall and winter months. Fatalities were more likely to occur in autumn and winter than in spring and summer (F: 15% VS 7%). This result is mainly because of the rugged terrain and mountainous roads in the Northwest. In addition, the harsh natural environment in autumn and winter means that vehicles and equipment are more likely to break down, thus increasing the likelihood of dangerous accidents [45].

#### 7) NORTHEAST CHINA

The following results can be observed from Figure 5 and Table 3. In Northeast China, FAF, SEA, DAF, HM, and MON are the key features, in order of importance, in distinguishing the severity of hazardous material road transport accidents.

Accidents where the final accident form was a spill accounted for 72% of the total number of accidents, and the fatality rate was significantly lower than that of other final accident forms (F: 8% VS 44%). The reasons for this result are similar to those described above for Central China.

Accidents occurring during the winter months accounted for 22% of the total number of accidents in the region. Contrary to previous perceptions, the probability of a fatal winter accident was extremely low and almost nonexistent, whereas the probability of a fatal accident in other seasons was 10%. This is probably because drivers understand the harshness of the winter environment in the Northeast, the difficulty of rescue, and the severity of an accident, so they increase their caution, thus reducing the chance of a serious accident.

Direct accident forms involving spills, fires, and two-vehicle rear-end collisions were less likely to result in fatalities than other direct accident forms (F: 3% VS 12%). Explanations for this result can be found in the sections on the North and Southwest regions of China. The direct form of an accident involved spills, fires and two-vehicle rear-end collisions in 49% of the accidents.

Among all hazardous material road transport accidents in the Northeast, 62% involved Class III hazardous materials, much higher than for other hazardous materials. This may also be one of the reasons why the probability of fatal accidents was higher for Class III hazardous materials than for other types of hazardous materials (F: 13% VS 0%).

Eleven percent of road transport accidents involving hazardous materials occurred in March. The fatal accident rate was extremely high compared to that occurring in other months (F: 29% VS 5%). This may be due to the fact that the hazardous material industry in the Northeast begins operations in March, and drivers are not fully familiar with the vehicles and routes to handle changes in the driving environment [3]. In addition, the excitement of starting work may cause drivers to forget the unique nature of hazardous material road transport, let their guard down, and engage in unsafe behaviors such as speeding.

### D. PROPOSALS TO IMPROVE SAFETY IN THE TRANSPORT OF HAZARDOUS MATERIALS BY ROAD

According to the results of the above analysis, corresponding recommendations will be made for each of the seven regions regarding how to improve the safety of hazardous material road transport.

#### 1) EAST CHINA

East China, with a relatively dense transportation network and population, should first establish relevant laws and regulations regarding safe distances for industries, while companies should try to avoid densely populated areas such as residential areas when choosing routes [9].

The safety of transporting Class III and VIII hazardous chemicals is strictly controlled, specific transport plans and workflows are formulated, safety education is carried out, and safety supervision of transport enterprises is strengthened [3].

Gather information from various sources (weather, road conditions) to adjust routes and transportation schedules in a timely manner to avoid driving in rain or snow or on slippery roads. Near curves, ramps, and other special road alignments, the road designer should provide sufficient information by installing road signs to alert drivers to upcoming road alignments and that they should reduce their speed and remain alert [42].

Truck manufacturers are recommended to take sufficient care in developing new safety equipment and detection instruments. For example, by adding driver detection devices to the in-vehicle system, the driver's driving time, mental state and operating behavior will be monitored in real time, and any unsafe behavior will be promptly alerted to reduce the occurrence of accidents or aggravation caused by fatigue [20].

Other options are additional mobile checkpoints for hazardous materials at suitable locations on roads and mandatory control of vehicle travel times.

Traffic control, such as drowsy driving checks, should be strengthened from 15:00-16:00.

#### 2) NORTH CHINA

Recommendations on the treatment of the FA, RA, and HM factors can be found in the section on the East China region.

Road authorities and transport companies should invest more in real-time monitoring and early warning equipment and establish monitoring and early warning systems. A coordination mechanism should be developed among transport enterprises, road authorities, fire departments, and environmental protection and health authorities to make efficient emergency rescues and plans [42].

Introduce standards for the hours of exclusion of vehicles transporting hazardous materials on urban roads, and strictly enforce them [48]. Speed limits should be strengthened on expressways and national, provincial roads and rural roads, and more road infrastructure should be installed for rural roads.

Reduce the design of longer, straighter roads, or add bulges in an orderly manner on longer, straighter roads to constantly remind drivers to stay alert [49].

The frequency of safety inspections of transport vehicles and equipment should be increased during the autumn and winter seasons. Additionally, traffic control should be strengthened. Recommendations on the transport of Class III hazardous chemicals can be found in the section on the East China region.

#### 3) CENTRAL CHINA

Recommendations for dealing with the DAF, FA, and RA factors can be found in the sections on the East China and North China regions.

Suggestions for handling the final form of an accident: Advanced technologies such as global positioning systems (GPS), geographic information systems (GIS), electronic billing and mobile networks can be used to set up monitoring systems to understand the development of accidents and provide basic information for the timely formulation of rescue plans [50].

Developing effective training programs for road transport accidents involving hazardous chemicals, raising staff risk awareness and knowledge of the characteristics of hazardous materials, and improving the response capacity are also necessary [10]. In addition, it is necessary to establish a linkage among transport enterprises, road management departments and emergency rescue organizations [50].

#### 4) SOUTH CHINA

For suggestions on the handling of the FAF, FA, RT, and SEA factors, please refer to the sections on East China, North China, and Central China.

The vehicle type should be regulated more heavily in terms of the overloading of tankers, and overweight vehicles have larger inertia and reduced operability. A load detection device can be installed at a load detection site for hazardous material transport vehicles, and the data can be uploaded in real time [51]. When a load is heavy, the relevant supervisors will be notified to avoid overloading in the transportation process, which can cause hazardous material transportation accidents.

The region will have to increase the costs of illegal modification and illegal transportation and improve the frequency and supervision of inspections on rural roads.

#### 5) SOUTHWEST CHINA

Recommendations for dealing with DAF and FA can be referenced in the section on the East China region.

In summer transportation, enterprises should comprehensively collect information from various parties, set up routes and schedules, avoid traveling on rainy days and steep terrain, and strengthen training for drivers to improve their safety awareness and ability to deal with emergencies.

Companies should tighten their load management of Class IV hazardous chemicals to avoid exposure to wet conditions during transport [52].

Recommendations for wet road surface conditions can be taken from the suggestions for summer transport management.

#### 6) NORTHWEST CHINA

Recommendations on the treatment of RT, FAF, FA, and DAF can be obtained from the sections on the relevant regions above.

Route planning should be undertaken cautiously to minimize driving on wet, waterlogged, icy and snowy roads. If necessary, additional vehicle anti-skid equipment can be installed.

In the autumn and winter seasons, transport enterprises need to perform proper vehicle maintenance to deal with the harsh natural environment and rugged terrain of the Northwest Territories. Additionally, the frequency of vehicle inspections should be increased to ensure that vehicles and equipment run well. Increasing the stockpile of emergency supplies and equipment is also necessary [15].

#### 7) NORTHEAST CHINA

Proposals for the DAF, FAF, and HM factors can be obtained from the sections on the abovementioned regions.

The Northeast region should strengthen the supervision of road transport of hazardous materials in spring, summer and autumn.

Transport companies should intensify training on driving skills, emergency operations and safety awareness before the resumption of work in March. Traffic authorities should increase the frequency of traffic control and inspections, such as by addressing speed limits, fatigue and illegal transport [12].

## V. CONCLUSION

This paper proposed the use of XGBoost to develop a ternary classification model of property damage only, injury, and fatal accidents. On the basis of this model, we explored the factors influencing the severity of hazardous material road transport accidents in seven regions of China. In addition, four popular models, LR, MLP, RPLM and SVM, were applied to model

the same data to validate the proposed model. XGBoost was found to have better prediction accuracy than the other models. It was then applied to explore the importance of factors influencing the severity of hazardous material road transport accidents in different areas, as well as to analyze the reasons why these important factors influence the severity of accidents in different regions. In the data, the distribution of hazardous material road transport accidents varied from region to region, and XGBoost performed well for those regions with a large amount of data (East China). Therefore, it is certain that more information is needed to obtain productive results.

The accident analysis results showed that there were some differences in the factors that determine the severity of hazardous material road transport accidents in different regions. The importance of the same factors in influencing the severity of accidents varied somewhat by region. There were also some regional differences in the causes of the impact of the same factor on the severity of accidents. Depending on the results of the analysis of the main influencing factors and causes identified in this study, targeted recommendations and countermeasures were provided for each region to improve the problems in the road transport of dangerous goods.

Nevertheless, this study is subject to several limitations. First, although this study collected data for 1411 hazardous material road transport accidents, the sample collected is relatively small compared to the research data of other traffic accidents. Second, due to the special nature of road transport accidents involving hazardous materials, the accident investigation cycle is relatively long. Some accidents were still under investigation at the time of accident data collection, and more comprehensive information was not available. Finally, the quality of the data is also limited by the professionalism of the collectors due to the lack of professionals in the study of hazardous material road transport in China. The sample size and dimension of the sample will be further expanded in future studies, and a more rational preprocessing approach to the data will be adopted to improve the quality of the data and perform a more comprehensive and prudent study.

## REFERENCES

[1] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," *Stata J., Promoting Commun. Statist. Stata*, vol. 6, no. 1, pp. 58–82, Feb. 2006, doi: 10.1177/1536867x0600600104.

[2] Y. Yuan, M. Yang, Z. Gan, J. Wu, C. Xu, and D. Lei, "Analysis of the risk factors affecting the size of fatal accidents involving trucks based on the structural equation model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 8, pp. 112–124, Aug. 2019.

[3] L. Zhao, Y. Qian, Q.-M. Hu, R. Jiang, M. Li, and X. Wang, "An analysis of hazardous chemical accidents in China between 2006 and 2017," *Sustainability*, vol. 10, no. 8, p. 2935, Aug. 2018, doi: 10.3390/su10082935.

[4] J. Current and S. Ratick, "A model to assess risk, equity and efficiency in facility location and transportation of hazardous materials," *Location Sci.*, vol. 3, no. 3, pp. 187–201, Oct. 1995, doi: 10.1016/0966-8349(95)00013-5.

[5] M. D. Abkowitz, J. P. DeLorenzo, R. Duych, A. Greenberg, and T. McSweeney, "Assessing the economic effect of incidents involving truck transport of hazardous materials," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1763, no. 1, pp. 125–129, Jan. 2001, doi: 10.3141/1763-18.

[6] B. Inanloo and B. Tansel, "A transportation network assessment tool for hazardous material cargo routing: Weighing exposure health risks, proximity to vulnerable areas, delay costs and trucking expenses," *J. Loss Prevention Process Industries*, vol. 40, pp. 266–276, Mar. 2016, doi: 10.1016/j.jlp.2016.01.002.

[7] S. E. Andersson, "Safe transport of dangerous goods: Road, rail or sea? A screening of technical and administrative factors," *Eur. J. Oper. Res.*, vol. 75, no. 3, pp. 499–507, 1994, doi: 10.1016/0377-2217(94)90292-5.

[8] J. Yang, F. Li, J. Zhou, L. Zhang, L. Huang, and J. Bi, "A survey on hazardous materials accidents during road transport in China from 2000 to 2008," *J. Hazardous Mater.*, vol. 184, nos. 1–3, pp. 647–653, Dec. 2010, doi: 10.1016/j.jhazmat.2010.08.085.

[9] H.-D. Zhang and X.-P. Zheng, "Characteristics of hazardous chemical accidents in China: A statistical investigation," *J. Loss Prevention Process Industries*, vol. 25, no. 4, pp. 686–693, Jul. 2012, doi: 10.1016/j.jlp.2012.03.001.

[10] X. Shen, Y. Yan, X. Li, C. Xie, and L. Wang, "Analysis on tank truck accidents involved in road hazardous materials transportation in China," *Traffic Injury Prevention*, vol. 15, no. 7, pp. 762–768, Oct. 2014, doi: 10.1080/15389588.2013.871711.

[11] C. Ma, J. Zhou, and D. Yang, "Causation analysis of hazardous material road transportation accidents based on the ordered logit regression model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 4, p. 1259, Feb. 2020, doi: 10.3390/ijerph17041259.

[12] W. Duan, G. Chen, Q. Ye, and Q. Chen, "The situation of hazardous chemical accidents in China between 2000 and 2006," *J. Hazardous Mater.*, vol. 186, nos. 2–3, pp. 1489–1494, Feb. 2011, doi: 10.1016/j.jhazmat.2010.12.029.

[13] M. Poku-Boansi, P. Tornyeviadzi, and K. K. Adarkwa, "Next to suffer: Population exposure risk to hazardous material transportation in Ghana," *J. Transp. Health*, vol. 10, pp. 203–212, Sep. 2018, doi: 10.1016/j.jth.2018.06.009.

[14] Y. Xing, S. Chen, S. Zhu, Y. Zhang, and J. Lu, "Exploring risk factors contributing to the severity of hazardous material transportation accidents in China," *Int. J. Environ. Res. Public Health*, vol. 17, no. 4, p. 1344, Feb. 2020.

[15] B. Fabiano, F. Curru, A. P. Reverberi, and R. Pastorino, "Dangerous good transportation by road: From risk analysis to emergency planning," *J. Loss Prevention Process Industries*, vol. 18, nos. 4–6, pp. 403–413, Jul. 2005, doi: 10.1016/j.jlp.2005.06.031.

[16] G. Azimi, A. Rahimi, H. Asgari, and X. Jin, "Severity analysis for large truck rollover crashes using a random parameter ordered logit model," *Accident Anal. Prevention*, vol. 135, Feb. 2020, Art. no. 105355, doi: 10.1016/j.aap.2019.105355.

[17] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Syst. Appl.*, vol. 91, pp. 452–463, Jan. 2018, doi: 10.1016/j.eswa.2017.09.025.

[18] F. Chen and S. Chen, "Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways," *Accident Anal. Prevention*, vol. 43, no. 5, pp. 1677–1688, Sep. 2011, doi: 10.1016/j.aap.2011.03.026.

[19] A. J. Khattak, D. Ph, and R. J. Schneider, "Risk factors in large truck rollovers and injury severity: Analysis of single-vehicle collisions," TRB Annu. Meet., Washington, DC, USA, Tech. Rep. TRB Paper: 03-2331, Jan. 2003.

[20] X. Li, T. Liu, and Y. Liu, "Cause analysis of unsafe behaviors in hazardous chemical accidents: Combined with HFACs and Bayesian network," *Int. J. Environ. Res. Public Health.*, vol. 17, no. 1, p. 11, Dec. 2019, doi: 10.3390/ijerph17010011.

[21] A. Tavakoli Kashani, R. Rabieyan, and M. M. Besharati, "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers," *J. Saf. Res.*, vol. 51, pp. 93–98, Dec. 2014, doi: 10.1016/j.jsr.2014.09.004.

[22] L. Zhao, X. Wang, and Y. Qian, "Analysis of factors that influence hazardous material transportation accidents based on Bayesian networks: A case study in China," *Saf. Sci.*, vol. 50, no. 4, pp. 1049–1055, Apr. 2012, doi: 10.1016/j.ssci.2011.12.003.

[23] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[24] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Anal. Prevention*, vol. 108, pp. 27–36, Nov. 2017, doi: 10.1016/j.aap.2017.08.008.

[25] G. Casalicchio, C. Molnar, and B. Bischl, "Machine learning and knowledge discovery in databases," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, in Lecture Notes in Computer Science, vol. 11051, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds. Cham, Switzerland: Springer, 2018, pp. 655–670, doi: 10.1007/978-3-030-10925-7_40.

[26] J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan, and J. Zhang, "Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: A city management perspective," *IEEE Access*, vol. 7, pp. 148059–148072, 2019.

[27] S. Soleimani, S. R. Mousa, J. Codjoe, and M. Leitner, "A comprehensive railroad-highway grade crossing consolidation model: A machine learning approach," *Accident Anal. Prevention*, vol. 128, pp. 65–77, Jul. 2019, doi: 10.1016/j.aap.2019.04.002.

[28] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405, doi: 10.1016/j.aap.2019.105405.

[29] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, and T. Li, "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Appl. Soft Comput.*, vol. 80, pp. 57–79, Jul. 2019, doi: 10.1016/j.asoc.2019.03.017.

[30] X. Shi, Q. Li, Y. Qi, T. Huang, and J. Li, "An accident prediction approach based on XGBoost," in *Proc. IEEE ISKE*, Nanjing, China, Nov. 2017, p. 7.

[31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794, 2016.

[32] C. Lin, D. Wu, H. Liu, X. Xia, and N. Bhattarai, "Factor identification and prediction for teen driver crash severity using machine learning: A case study," *Appl. Sci.*, vol. 10, no. 5, p. 1675, Mar. 2020, doi: 10.3390/app10051675.

[33] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning anders," presented at the Natural Synth. NIPS, Denver, CO, USA, Nov./Dec. 1994.

[34] D. M. W. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, Dec. 2007. [Online]. Available: http://www.bioinfo.in/contents.php?id=51

[35] S. Mafi, Y. A. Razig, and R. Doczy, "Machine learning methods to analyze injury severity of drivers from different age and gender groups," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 38, pp. 171–183, Dec. 2018, doi: 10.1177/0361198118794292.

[36] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[37] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," in *Proc. World Enformatika Soc.*, Vienna, Austria, Mar. 2006, pp. 278–283.

[38] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked science in machine learning," *ACM SIGKDD Explor. Newsletter*, vol. 15, no. 2, pp. 49–60, 2014, doi: 10.1145/2641190.2641198.

[39] I. Radun and J. E. Radun, "Convicted of fatigued driving: Who, why and how?" *Accident Anal. Prevention*, vol. 41, no. 4, pp. 869–875, Jul. 2009, doi: 10.1016/j.aap.2009.04.024.

[40] W. Zou, X. Wang, and D. Zhang, "Truck crash severity in New York city: An investigation of the spatial and the time of day effects," *Accident Anal. Prevention*, vol. 99, pp. 249–261, Feb. 2017, doi: 10.1016/j.aap.2016.11.024.

[41] G. Zhang, K. K. W. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Anal. Prevention*, vol. 87, pp. 34–42, Feb. 2016, doi: 10.1016/j.aap.2015.10.033.

[42] D. M. Goldberg and S. Hong, "Minimizing the risks of highway transport of hazardous materials," *Sustainability*, vol. 11, no. 22, p. 6300, Nov. 2019, doi: 10.3390/su11226300.

[43] Y. Chen, K. Wang, Y. Zhang, and Q. Shi, "Identification of black spots on highways using fault tree analysis and vehicle safety boundaries," *J. Transp. Saf. Secur.*, pp. 1–23, 2019, doi: 10.1080/19439962.2019.1605639.

[44] J. Hong, R. Tamakloe, and D. Park, "Application of association rules mining algorithm for hazardous materials transportation crashes on expressway," *Accident Anal. Prevention*, vol. 142, Jul. 2020, Art. no. 105497, doi: 10.1016/j.aap.2020.105497.

[45] J. P. Thompson, M. R. J. Baldock, J. L. Mathias, and L. N. Wundersitz, "An examination of the environmental, driver and vehicle factors associated with the serious and fatal crashes of older rural drivers," *Accident Anal. Prevention*, vol. 50, pp. 768–775, Jan. 2013, doi: 10.1016/j.aap.2012.06.028.

[46] J.-L. Song and D.-H. Wang, "Tracking laser Doppler measurement for velocity of moving target," in *Proc. Int. Conf. Comput. Sci. Inf. Process. (CSIP)*, vol. 32, Aug. 2012, pp. 426–431, doi: 10.1109/CSIP.2012.6308884.

[47] A. Iranitalab, Y. Kang, and A. Khattak, "Modeling the probability of hazardous materials release in crashes at highway—Rail grade crossings," *Transp. Res. Rec.*, vol. 2672, no. 10, pp. 28–37, 2018, doi: 10.1177/0361198118780885.

[48] C. Caliendo and M. L. D. Guglielmo, "Quantitative risk analysis on the transport of dangerous goods through a bi-directional road tunnel," *Risk Anal.*, vol. 37, no. 1, pp. 116–129, Jan. 2017, doi: 10.1111/risa.12594.

[49] T. Yared, P. Patterson, and E. S. A. All, "Are safety and performance affected by navigation system display size, environmental illumination, and gender when driving in both urban and rural areas?" *Accident Anal. Prevention*, vol. 142, Jul. 2020, Art. no. 105585, doi: 10.1016/j.aap.2020.105585.

[50] W. H. Tate and M. D. Abkowitz, "Emerging technologies applicable to hazardous materials transportation safety and security," *J. Transp. Saf. Secur.*, vol. 4, no. 3, pp. 244–257, Sep. 2012, doi: 10.1080/19439962.2012.657286.

[51] X. Liu and C. T. Dick, "Risk-based optimization of rail defect inspection frequency for petroleum crude oil transportation," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2545, no. 1, pp. 27–35, Jan. 2016, doi: 10.3141/2545-04.

[52] H. Sung, J. Kim, J. Hong, D. Park, and Y.-I. Lee, "Transport management characteristics of urban hazardous material handling business entities," *Sustainability*, vol. 11, no. 23, p. 6600, Nov. 2019.

**XIAOYAN SHEN** received the Ph.D. degree in transportation engineering from Chang'an University, Xi'an, China, in 2009. She is currently an Associate Professor with the School of Automobile, Chang'an University, and also a Master's Tutor. Her main research interests include road transportation safety technology, road dangerous goods' transportation safety management, and highway service area operation management.

**SHANSHAN WEI** received the B.S. degree in transportation engineering from the Nanjing Forest University, Nanjing, China. She is currently pursuing the master's degree in engineering with the School of Automobile, Chang'an University, Xi'an, China.