Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 1570521, 14 pages
https://doi.org/10.1155/2022/1570521

*Research Article*

# A Roadway Safety Sustainable Approach: Modeling for Real-Time Traffic Crash with Limited Data and Its Reliability Verification

**Zhenzhou Yuan**,[1] **Kun He**,[1] **and Yang Yang**[1,2]

[1]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Transportation Science and Engineering, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Yang Yang; 14114222@bjtu.edu.cn

Received 1 December 2021; Revised 19 December 2021; Accepted 27 December 2021; Published 15 January 2022

Academic Editor: Wenxiang Li

Copyright © 2022 Zhenzhou Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of freeway system informatization, it is easier to obtain the traffic flow data of freeway, which are widely used to study the relationship between traffic flow state and traffic safety. However, as the development degree of the freeway system is different in different regions, the sample size of traffic data collected in some regions is insufficient, and the precision of data is relatively low. In order to study the influence of limited data on the real-time freeway traffic crash risk modeling, three data sets including high precision data, small sample data, and low precision data were considered. Firstly, Bayesian Logistic regression was used to identify and predict the risk of three data sets. Secondly, based on the Bayesian updating method, the migration test towards high and low precision data sets was established. Finally, the applicability of machine learning and statistical methods to low precision data set was compared. The results show that the prediction performance of Bayesian Logistic regression improves with the increasing of sample size. Bayesian Logistic regression can identify various significant risk factors when data sets are of different precision. Comparatively, the prediction performance of the support vector machine is better than that of Bayesian Logistic. In addition, Bayesian updating method can improve the prediction performance of the transplanted model.

## 1. Introduction

In recent years, the potential safety hazard of new energy vehicles has gradually attracted attention, especially the accident of pure electric vehicles [1, 2]. As new energy vehicles and shared vehicles enter the freeway, they are also facing the risk of traffic collisions. As one of the important subsystems of the road system, the freeway greatly facilitates people's travel and improves the transportation efficiency of goods. At the same time, because of the large traffic volume and fast speed of vehicles on freeways, relatively serious traffic crashes are easy to occur, which brings great harm to the safety of people's lives and property. Freeway traffic crash has become one of the problems that cannot be ignored [3].

A large number of scholars have done extensive researches on traffic safety. Some scholars analyzed the internal relationship between the factors causing accidents and the distribution law of accidents based on the historical traffic accident data collected and then put forward the corresponding countermeasures. For example, considering the differences in time, Yuan et al. adopted an improved association rule mining algorithm to analyze the association among the influencing factors of freeway traffic crashes, in which the hidden association rules were found and the accuracy of the algorithm was improved [4]. Tian et al. analyzed the temporal and spatial distribution characteristics of freeway crashes in mountainous areas based on historical crash data, identified the significant influencing factors, and proposed corresponding improvement strategies [5]. Some scholars have analyzed the main influencing factors of accidents according to specific accident scenarios. Mergia et al. analyzed the crash at the junction. It was pointed out that drunk driving and overspeed have increased the severity of crashes in the diverging area, bad weather increased the severity of crashes in the merging area, and adverse linear conditions

would increase the severity of crashes in the diverging area [6]. Xin et al. studied the factors not observed, proving that, under different severities, driving behaviors, environmental characteristics, and other factors have significant differences in the impact of the crash rate [7]. Haghighi et al. studied the impact of road design features on crashes and found that 10-foot wide lanes and narrower shoulder were significantly associated with crash severity and increased vehicle density and guardrail length could reduce crash severity [8]. In order to study the influence of drivers' ages on crash severity, Osman et al. constructed a generalized ordered response probit model to reduce the interference of heterogeneity and found that each variable in different age groups had a different influence on crash [9]. Xu et al. introduced the Bayesian spatial random coefficient model to consider the heterogeneity of spatial structure and unstructured data when studying the spatial variation law of crash rate and cause factors, which improved the fitting effect of the model and verified that the existence of spatial structure heterogeneity would cause bias to parameter estimation [10]. Wang et al. explored risk factors' influence on urban traffic crashes frequency while considering both the spatial and temporal correlation/heterogeneity of traffic crashes. The linear regression model, spatial lag model (SLM), spatial error model (SEM), and time-fixed effects error model (T-FEEM) were established and compared, respectively [11]. To figure out the factors relating to crash risk in different regional types and their inner relation, Yang et al. took three sections of highway (areas of downtown, suburb, and mountain, in Washington State, USA) as the research object and, based on AHP improved Apriori association rule mining algorithm, identified the crash risk influencing factors and their complex association rules were [12]. Li et al. investigated the possibility of using support vector machine (SVM) models for crash injury severity analysis and compared the performance of the SVM model and the order probit model. It was found that the SVM model produced better prediction performance for crash injury severity than did the OP model [13]. In addition, Logit and Tobit models are also widely used in traffic crash analysis [14–22].

With the development of freeway informatization and the improvement of dynamic traffic management, the real-time crash risk model has been widely studied [23–26]. Based on loop detector data and crash data collected by the Shanghai expressway system, Sun et al. established a Bayesian network (BN) model to analyze real-time traffic flow parameters and crash risk of expressway [27]. You et al. established a support vector machine model to analyze highway traffic flow data for rear-end crash. The results showed that the SVM classifier has high practical value and reliability of real-time crash prediction based on traffic flow data of a single volume detector [28]. Xu et al. established a crash risk prediction model based on traffic flow data and meteorological data by using the Logistic model based on American freeway data. The results showed that weather conditions have a significant impact on crash risk [29]. Ma et al. established a crash risk assessment and analysis model using highway crash data and real-time traffic flow data. The significant variables were selected by a random forest algorithm, and the support vector machine model was established. The evaluation ability of models under different kernel functions was compared. The results showed that the model could effectively evaluate road crash risk based on real-time traffic flow [30, 31].

The traditional "postevent" traffic safety analysis can analyze the main influencing factors of crash occurrence, but it is difficult to reflect the influence of dynamic traffic flow characteristics on crash risk. At present, most of the researches on using traffic flow data to establish real-time crash risk models are based on existing data for modeling and analysis. However, different regions have different levels of development, and the data collection of traffic flow and traffic crash will be different. Then does the traffic flow variable also have a significant impact on the occurrence of traffic crashes? If the impact is small, can certain technical means be used to improve the accuracy of the corresponding model?

In order to study the above questions, based on the basic data necessary for the real-time crash risk model, this paper constructs three types of data sets: (1) high precision data set; (2) small sample data set; (3) low precision data set. On the basis of the above three types of data sets, statistical and machine learning methods are used to study the classification and prediction performance of the model under different data sets, and the applicability of the two methods is further analyzed.

From the perspective of data, this paper uses statistical Logistic regression, Bayesian theory, and support vector machine to simulate the impact of different types of data on real-time crash risk modeling. Furthermore, the applicability of different methods under different data types is compared. The conclusions of this paper can be used as a reference for the subsequent practice and research of highway traffic safety.

## 2. Data Description

*2.1. The Data Source.* This paper selects the traffic flow and traffic crash data of milepost 100–132 of I-5 in Washington State in 2016. Figure 1 describes the main freeway section in the study area.

In 2016, a total of 332 traffic crashes occurred in this freeway section. In the selected area, 152 groups of loop detectors are arranged bidirectional, and the average

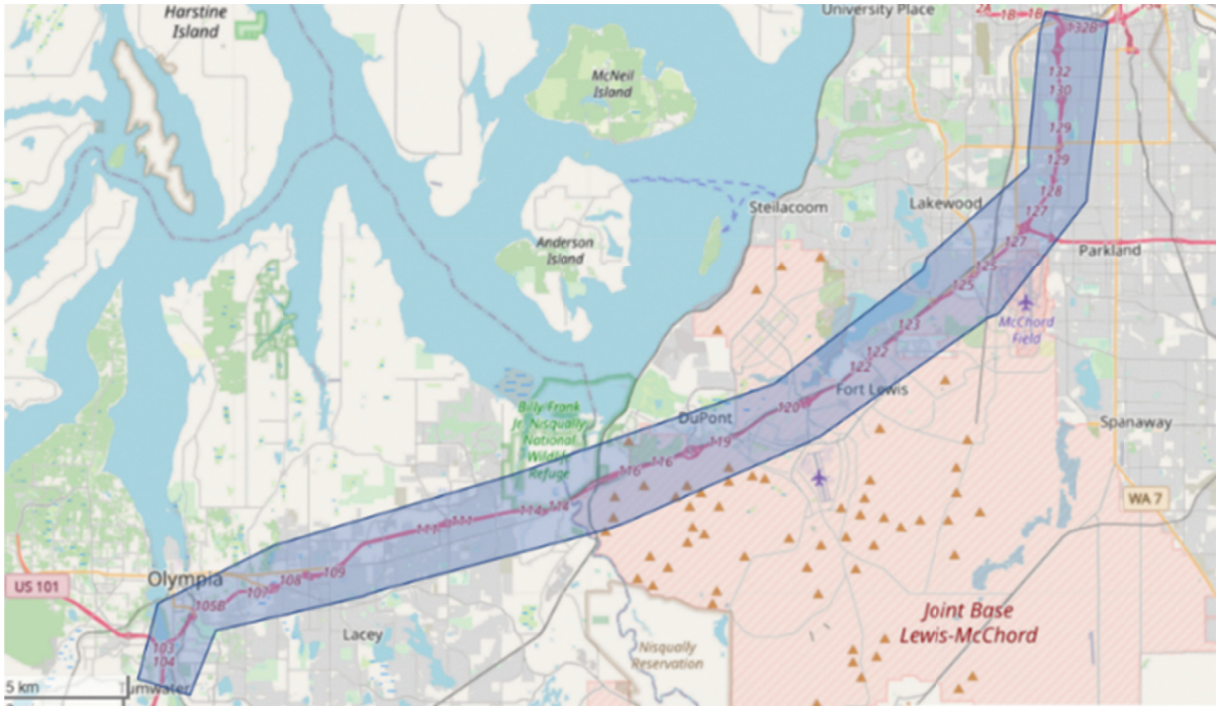


Figure 1: Study section of freeway.

distance between adjacent loops is about 0.7 km. Each loop detector collects average speed, occupancy, and traffic volume in each lane over a 20-second period.

*2.2. Variable.* In existing studies, traffic flow data of 5-minute lumps are used for analysis, and good research results are obtained [8–13]. Therefore, this paper adopts traffic flow data of 5-minute lumps for analysis, mainly including the volume, speed, and occupancy rate of each lane. The time period of 5–10 minutes before the crash is selected and two groups of upstream and downstream loop detectors were taken into account, as shown in Figure 2.

In addition to the basic data detected by the loop detector above, such as volume, speed, and occupancy of upstream and downstream loops, this paper combines the traffic flow variables as follows [2]. Considering that the difference values in volume, speed, and occupancy between upstream and downstream loops may lead to vehicle crash, the absolute value of the difference values between volume, speed, and occupancy between upstream and downstream loops is constructed. At the same time, the lateral crash between lanes is also one of the main forms of crash. The average difference values of volume, speed, and occupancy between adjacent lanes are constructed to describe the related variables of lateral collision. The specific meanings are shown in Table 1.

*2.3. Sample Structure Design*

*2.3.1. High Precision Data Set Sample.* High precision data set refers to the traffic flow data and traffic crash data collected by the American freeway system as the standard. The freeway loop

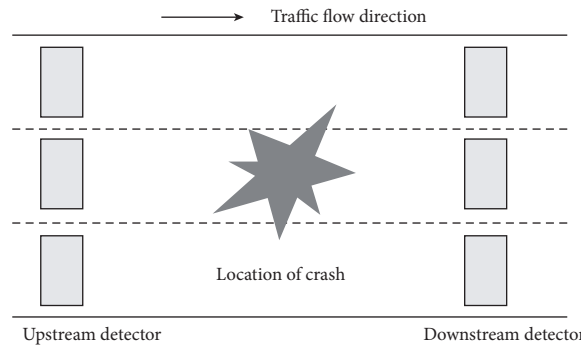


Figure 2: Traffic crash diagram.

detector in the United States has a high laying density, and the traffic crash information is collected completely.

In this paper, the paired sampling method is adopted to match control samples, and noncrash data under the same conditions are extracted for each crash data with a ratio of 1 : 4. The ratio of crash data to noncrash data is 1 : 4 for matching [2]. After data preprocessing, 191 traffic crash data and 764 noncrash data are obtained. The high precision data set sample is shown in Figure 3.

*2.3.2. Small Sample Data Set.* In order to study the influence of data sample size on the real-time crash risk model constructed, it is necessary to obtain small sample data with different sample sizes. The main ideas of constructing small sample data set in this paper are as follows: obtain and match the high precision data set, and extract data from the high precision data set in a proportion of 5%, 10%, 20%, 30%, and 50%, so as to construct small sample data set of different proportions. The small sample data set is shown in Figure 4.

Table 1: Variable name and physical meaning.

| Variable | Meaning | Variable | Meaning |
|---|---|---|---|
| *up_v* | The upstream volume | *down_o* | Average downstream occupancy |
| *up_s* | Average speed of upstream traffic volume | *down_dif_v* | Average absolute value of volume difference between adjacent downstream lanes |
| *up_o* | Average upstream occupancy | *down_dif_s* | Average absolute value of speed difference between adjacent downstream lanes |
| *up_dif_v* | Average absolute value of volume difference between adjacent upstream lanes | *down_dif_o* | Average absolute value of occupancy difference between downstream adjacent lanes |
| *up_dif_s* | Average absolute value of speed difference between adjacent upstream lanes | *abs_dif_v* | Absolute value difference between upstream and downstream volume |
| *up_dif_o* | Average absolute value of occupancy difference between upstream adjacent lanes | *abs_dif_s* | Absolute value of speed difference between upstream and downstream |
| *down_v* | The downstream volume | *abs_dif_o* | Absolute value difference between upstream and downstream occupancy |
| *down_s* | Average speed of downstream traffic volume | | |

| up_v | up_s | up_o | up_dif_v | up_dif_s | up_dif_o | down_v | down_s | down_o | own_dif | own_dif | own_dif | bs_dif_ | bs_dif_ | bs_dif_ | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| −0. 68441 | 0. 163659 | −0. 29517 | −0. 32964 | 0. 329017 | −0. 72624 | −0. 36068 | 0. 546614 | −0. 16984 | −0. 05247 | 0. 682967 | −0. 54752 | −0. 68043 | −0. 3625 | −0. 68998 | 1 |
| 0. 421289 | −0. 77021 | 1. 715074 | −1. 41983 | −0. 25792 | −0. 73558 | −0. 09158 | −2. 13843 | 3. 059773 | −0. 80973 | 0. 149726 | −0. 53251 | −0. 06695 | 1. 626693 | 0. 505555 | 1 |
| −0. 51135 | −0. 50202 | 0. 255297 | −0. 69304 | −0. 63884 | 0. 059538 | 1. 020696 | 0. 123974 | 1. 05477 | 0. 031674 | −1. 38998 | 0. 574897 | 1. 35454 | 0. 032906 | 0. 050371 | 1 |
| 0. 652044 | −0. 69696 | −0. 16167 | 0. 985514 | 1. 47038 | 0. 245664 | 0. 626018 | 0. 006519 | −0. 0475 | −1. 07618 | 0. 390673 | −0. 24468 | −0. 90488 | 0. 159818 | −0. 72195 | 1 |
| 1. 719285 | −2. 41331 | 3. 127089 | −0. 918 | 0. 383031 | 5. 506444 | −1. 5178 | −1. 13514 | −1. 07261 | −0. 09454 | 1. 33989 | −0. 71158 | 4. 332187 | 1. 087805 | 4. 979543 | 1 |
| 0. 450133 | −2. 71424 | 2. 65425 | −1. 09104 | −0. 92286 | 1. 530871 | −1. 43707 | −1. 93982 | −0. 11723 | 0. 578587 | −0. 68959 | 0. 309929 | 2. 222397 | 0. 185483 | 3. 179099 | 1 |
| 1. 738515 | −2. 19604 | 2. 348171 | −0. 88339 | 0. 216537 | 2. 587597 | −0. 87196 | 0. 230103 | −0. 7332 | 1. 335851 | −0. 01299 | −0. 21125 | 3. 284774 | 3. 120983 | 3. 511914 | 1 |
| 1. 103939 | −1. 20243 | 1. 242066 | −0. 81417 | −0. 42792 | 1. 290966 | 0. 868207 | −0. 5683 | −0. 16987 | 0. 774915 | −0. 49772 | −0. 11744 | −0. 60562 | 0. 012769 | 1. 329067 | 1 |
| 0. 969332 | −0. 28245 | −0. 17012 | 0. 1895 | 2. 10045 | −0. 12562 | 1. 110395 | 0. 63278 | −0. 20174 | −1. 13227 | −0. 62691 | −0. 36594 | −0. 80014 | 0. 553284 | −0. 54618 | 1 |
| −0. 29982 | −0. 89755 | 0. 894417 | 1. 123951 | 0. 253046 | 0. 489311 | −1. 00651 | −2. 61024 | 4. 226433 | −0. 47317 | −0. 08751 | 0. 227479 | 0. 337053 | 2. 239837 | 3. 033843 | 1 |
| −1. 65551 | 0. 110034 | −0. 97014 | 0. 224109 | 0. 215935 | −0. 31611 | −0. 44141 | −0. 05725 | −0. 47678 | −1. 21641 | 0. 415572 | −1. 00761 | 0. 696166 | −0. 53245 | −0. 14774 | 1 |
| 0. 104001 | −2. 25757 | 2. 774571 | 0. 1895 | −0. 1591 | 3. 804259 | 1. 594773 | −1. 9702 | 1. 927871 | −0. 57133 | −1. 01465 | 0. 292557 | 1. 35454 | −0. 65011 | 0. 859884 | 1 |
| −0. 43443 | −0. 1565 | −0. 07879 | −0. 01816 | 0. 186951 | −0. 62702 | −0. 25304 | −0. 95276 | 0. 696943 | −0. 80973 | −0. 40417 | −0. 74973 | −0. 88992 | 0. 588801 | 0. 068833 | 1 |
| −1. 03054 | −1. 50621 | 0. 592504 | −0. 43347 | 0. 25027 | 0. 527696 | −0. 46832 | −1. 04277 | 0. 795404 | −1. 02008 | 0. 037817 | 0. 186618 | −0. 32132 | −0. 30291 | −0. 72145 | 1 |
| −0. 1556 | −1. 82467 | 1. 684171 | −0. 84878 | 0. 567499 | −0. 1299 | −0. 13643 | −1. 82607 | 1. 919292 | −0. 72559 | 0. 226412 | 0. 231857 | −0. 88992 | −0. 72944 | −0. 60752 | 1 |
| 0. 200149 | 0. 756686 | −0. 12365 | 0. 812467 | 0. 889828 | −0. 71253 | 0. 17752 | 1. 109791 | −0. 09759 | 0. 031674 | 0. 169834 | −0. 85501 | −0. 85999 | −0. 38589 | −0. 60961 | 1 |
| −0. 96324 | −0. 46531 | −0. 31264 | −1. 10835 | 2. 544329 | −0. 77833 | −0. 57596 | 0. 275311 | −0. 22955 | −1. 23043 | 0. 439101 | −0. 96425 | −0. 60562 | 0. 236663 | −0. 70559 | 1 |
| 1. 219316 | 0. 22487 | 0. 077173 | −0. 15659 | 0. 004811 | 0. 239609 | 1. 379494 | 0. 00871 | 0. 565516 | 0. 073744 | 1. 735347 | 0. 056289 | −0. 74029 | −0. 45199 | −0. 30206 | 1 |
| −0. 61711 | 0. 497055 | −0. 75549 | −0. 84878 | −0. 15521 | −0. 50233 | 0. 168551 | 0. 430877 | −0. 39905 | −1. 24446 | −0. 74949 | −0. 6145 | 0. 097644 | −0. 72963 | −0. 34432 | 1 |
| 1. 48853 | −0. 86834 | 0. 931766 | −0. 1912 | −0. 52749 | 1. 148165 | 1. 433313 | −0. 81331 | 0. 34867 | −1. 10422 | 0. 243402 | 0. 104932 | −0. 94977 | −0. 87595 | 0. 279127 | 1 |
| −1. 51128 | −3. 53585 | 6. 291478 | −1. 78323 | −1. 6648 | 0. 774712 | −1. 22179 | −2. 48337 | 0. 723412 | 0. 915149 | −0. 84303 | 2. 572265 | −0. 83007 | 0. 634922 | 7. 088264 | 1 |

Figure 3: High precision data set sample.

| up_v | up_s | up_o | up_dif_v | up_dif_s | up_dif_o | down_v | down_s | down_o | own_dif | own_dif | own_dif | bs_dif_ | bs_dif_ | bs_dif_ | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0. 402059 | −1. 81665 | 1. 619915 | 2. 231449 | 0. 799052 | 3. 621474 | 1. 191125 | −0. 96251 | 0. 504233 | 2. 822334 | 0. 258456 | 2. 641339 | 0. 217348 | 0. 370138 | 1. 022954 | 1 |
| −0. 35751 | −0. 87518 | 0. 012122 | −1. 57558 | 0. 680598 | −0. 18559 | −0. 16334 | −0. 80451 | −0. 30894 | 0. 01765 | −0. 30431 | −0. 64557 | −0. 85999 | −0. 90317 | −0. 16908 | 0 |
| 0. 027082 | 0. 702722 | −0. 46017 | −0. 08737 | 0. 647467 | 0. 054882 | 0. 3031 | 0. 18034 | −0. 2712 | −0. 79571 | 0. 300822 | −0. 35774 | −0. 68043 | 0. 06398 | −0. 58938 | 0 |
| 0. 690503 | −0. 0901 | 0. 096558 | 0. 656725 | −0. 09514 | −0. 33982 | 1. 98048 | 0. 594676 | −0. 55407 | −0. 4311 | 0. 504736 | −0. 52743 | 1. 085205 | 0. 156749 | 0. 242889 | 0 |
| 2. 363476 | −1. 51613 | 1. 27373 | −0. 57191 | −0. 12078 | 0. 57125 | −2. 12776 | −3. 29392 | 7. 387506 | −1. 02008 | −1. 42513 | 3. 3015 | 6. 352199 | 2. 384989 | 6. 356577 | 0 |
| −0. 64595 | −0. 44596 | −0. 07684 | −1. 62749 | 0. 958155 | −0. 29038 | 0. 26722 | −0. 97947 | 1. 138365 | −1. 10422 | 1. 596822 | −0. 47518 | 0. 307126 | 0. 140141 | 0. 601986 | 1 |
| −1. 57859 | 0. 54386 | −0. 9232 | −0. 01816 | −1. 38892 | −0. 60037 | −1. 32943 | 1. 009968 | −0. 87773 | 2. 219327 | −0. 58681 | 0. 113653 | −0. 90488 | −0. 19729 | −0. 69803 | 0 |
| 0. 267452 | 0. 172516 | −0. 89415 | 0. 293327 | −0. 33743 | −0. 30269 | −0. 36068 | 0. 39668 | −0. 26415 | 1. 08343 | 0. 514419 | 0. 646959 | 0. 142533 | −0. 64182 | 0. 007344 | 0 |
| −1. 02093 | 0. 941156 | −0. 66739 | −0. 58921 | 1. 571614 | −0. 87815 | −0. 71947 | 1. 043031 | −0. 59838 | −0. 51524 | −1. 03445 | −0. 83104 | −0. 75525 | −0. 81942 | −0. 70566 | 0 |
| 1. 084709 | −1. 3259 | 1. 153409 | −0. 36425 | −0. 38715 | −0. 13133 | 1. 845931 | −0. 76926 | 0. 673939 | 0. 129838 | −0. 10407 | 0. 41764 | 0. 247274 | −0. 12984 | 0. 184718 | 0 |
| −0. 61711 | 0. 479366 | −0. 50921 | 0. 085672 | −0. 89762 | −0. 64201 | −0. 37862 | 0. 619881 | −0. 39593 | −0. 26282 | −0. 45781 | −1. 01116 | −0. 8151 | −0. 77408 | −0. 67431 | 0 |
| −0. 00176 | 0. 415662 | 0. 112207 | 0. 241414 | 2. 523276 | 0. 055939 | 0. 814387 | 0. 692801 | −0. 28082 | −0. 34696 | 3. 318966 | −0. 50381 | 0. 217348 | −0. 53652 | −0. 06756 | 0 |
| −0. 45366 | 0. 961781 | −0. 7281 | −1. 00452 | −0. 42489 | −0. 75109 | 0. 016061 | −0. 44247 | −0. 38436 | 0. 354212 | −0. 55563 | −0. 24947 | −0. 4111 | 1. 604886 | −0. 36362 | 0 |
| 0. 238608 | 0. 48221 | −0. 28059 | −0. 90069 | 0. 490578 | −0. 67397 | 0. 096791 | −0. 72272 | −0. 19109 | 1. 223664 | −0. 01875 | 0. 348149 | −0. 66547 | 1. 277339 | −0. 70883 | 1 |
| −1. 67473 | −0. 27083 | −0. 91104 | −0. 90069 | 0. 256389 | −0. 45977 | −1. 03342 | −0. 6536 | −1. 18339 | 0. 213978 | 0. 654404 | −0. 7359 | −0. 26147 | −0. 13406 | −0. 35885 | 0 |
| 1. 123168 | −0. 74023 | 0. 0853 | −0. 01816 | 0. 115887 | 0. 302784 | −0. 74638 | −0. 22387 | −0. 55407 | 0. 494447 | 0. 498116 | −0. 46142 | 2. 117656 | −0. 17197 | 0. 227631 | 0 |
| −1. 03054 | −1. 19417 | −0. 17364 | −0. 84878 | −1. 17074 | −0. 79538 | −1. 08724 | −1. 52766 | 0. 102975 | 0. 031674 | 1. 122836 | 0. 38967 | −0. 66547 | −0. 17543 | −0. 52356 | 0 |
| 0. 700118 | 0. 380616 | −0. 30148 | 0. 743249 | 0. 151765 | −0. 01418 | 0. 25825 | 0. 816843 | 0. 310958 | −0. 59938 | 1. 461072 | 0. 070187 | −0. 21658 | −0. 25797 | −0. 09785 | 0 |
| 0. 82511 | 0. 077623 | 0. 247838 | 2. 577542 | 0. 543329 | 0. 502706 | 1. 218035 | −0. 26948 | −0. 08738 | 2. 85038 | −0. 69332 | 0. 730347 | −0. 39614 | −0. 21406 | −0. 11853 | 0 |
| −1. 28053 | 0. 966189 | −0. 73091 | 0. 241414 | −1. 21338 | −0. 53491 | −1. 55368 | 0. 887645 | −0. 95545 | 0. 368236 | −1. 19342 | −0. 6752 | −0. 27643 | −0. 73094 | −0. 39139 | 0 |
| 0. 834725 | 0. 390852 | −0. 2818 | 0. 500983 | −0. 33818 | −0. 48605 | 0. 599108 | 0. 089139 | −0. 32308 | −0. 0104 | 0. 03159 | −0. 57608 | −0. 57569 | −0. 30944 | −0. 55027 | 0 |

Figure 4: Small sample data set.

*2.3.3. Low Precision Data Set Sample.* Low precision data set refers to the data set constructed from the data collected by the detectors with lower density compared to the US freeway system. Considering that the data is difficult to obtain, this paper constructs a low precision data set through certain manual processing methods. Compared with the freeway system in the United States, many freeways in China do not have complete loop detection devices, and the distance between the detectors is relatively long. In this paper, the average distance between detectors of a certain section of freeway in China is taken as a reference, and the freeway data of the US is used to construct a low precision data set. The low precision data set sample is shown in Figure 5.

The main processing ideas are shown in Figure 6. Manually delete part of the loop number in the loop file so that the average distance between the remaining loops is approximately equal to the reference value. Then the processed loop file is used to match the data set to get the low precision data set. After screening, there are a total of 32 bidirectional loops. The low precision data set is screened by paired sampling method, and 161 crash data and 644 noncrash data were obtained.

## 3. Real-Time Crash Risk Prediction Model

*3.1. Bayesian Logistic Regression.* Logistic regression is a generalized linear regression model commonly used in statistical methods. Based on the binomial Logistic regression model, this paper establishes a crash risk model between freeway crashes and real-time traffic flow [23, 28]. The crash probability corresponding to a certain data in the research data set is shown as follows:

$$P(x_i) = \frac{1}{1 + e^{-x_i'\beta}}, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $x_i$ represents the $i$ th data; $P(x_i)$ represents the probability value of crash occurrence; $-x_i'\beta$ represents a linear combination of explanatory variables and their coefficients.

The Bayesian method is used to estimate the coefficients of the Logistic regression model. The Bayesian method assumes that all unknown parameters in the model are random variables. Before establishing the Bayesian model $M$, it is necessary to set the prior probability distribution $\pi(\Theta \mid M)$ of all parameters $\Theta$ in the model, which represents the known information of this parameter before obtaining the training data $Y$. After obtaining the training data $Y$, the Bayesian statistical model makes statistical inference on $\Theta$ through the posterior probability distribution. According to the Bayesian theorem, the posterior probability distribution $f(\Theta \mid Y, M)$ of parameter $\Theta$ in model $M$ can be expressed as follows:

$$f(\Theta \mid Y, M) = \frac{f(Y, \Theta \mid M)}{f(Y \mid M)} = \frac{f(Y \mid \Theta, M)\pi(\Theta \mid M)}{\int f(Y, \Theta \mid M) d\Theta} \propto f(Y \mid \Theta, M)\pi(\Theta \mid M). \tag{2}$$

Formula (2) shows that the posterior probability distribution of parameter $\Theta$ takes into account both the information contained in the training data $Y$ and the known information of parameter $\Theta$. $f(\Theta \mid Y, M)$ is the posterior distribution of parameter $\Theta$ in model $M$ under given training data $Y$. $f(Y, \Theta \mid M)$ is the joint probability distribution of $Y$ and $\Theta$ in model $M$. $f(Y \mid M)$ represents the marginal probability distribution of model $M$, that is, the probability distribution of training data $Y$ under given conditions. $\pi(\Theta \mid M)$ represents the prior probability distribution of parameter $\Theta$ in model $M$ before obtaining the training data $Y$. $f(Y \mid \Theta, M)$ is the likelihood function of model $M$.

*3.2. Bayesian Updating Method.* Based on the Bayesian updating method, the Bayesian Logistic regression model is established to transmigrate the real-time crash of freeways [32]. The Bayesian method can obtain the posterior probability distribution of each parameter in the model so that the prior probability distribution can be reset during model transplantation.

That is, when low precision data set $Y_1$ is used to establish a real-time crash risk model, the Bayesian method can be used to obtain the posterior probability distribution of each risk factor. When high precision data set $Y_2$ needs to establish the Logistic regression model and transplant it to low precision data set, the posterior probability distribution of the risk factors in the previous model can be used as the prior probability distribution of the risk factors in the new model, as shown in the following formula:

$$\begin{aligned} \pi(\beta \mid Y_1, Y_2) &\propto f(Y_1, Y_2 \mid \beta)\pi(\beta) \\ &= f(Y_2 \mid Y_1, \beta) f(Y_1 \mid \beta)\pi(\beta) \propto f(Y_2 \mid \beta)\pi(\beta \mid Y_1). \end{aligned} \tag{3}$$

Schematic diagram of the Bayesian updating method is shown in Figure 7.

$\pi(\beta \mid Y_1, Y_2)$ is the posterior distribution of parameter $\beta$ under given data sets $Y_1$ and $Y_2$; $f(Y_1, Y_2 \mid \beta)$ is the likelihood function; $\pi(\beta)$ is the prior probability distribution of parameter $\beta$; $f(Y_2 \mid Y_1, \beta)$ is the likelihood function given data set $Y_1$ and parameter $\beta$; $f(Y_1 \mid \beta)$ and $f(Y_2 \mid \beta)$ are the likelihood functions; $\pi(\beta \mid Y_1)$ is the posterior distribution of parameter $\beta$ under given data set $Y_1$.

*3.3. Support Vector Machine.* Support vector machine (SVM) is a machine learning classification algorithm based on statistical theory [26, 27]. It can obtain the optimal solution through existing information and can deal with small samples or limited samples well. In the sample space, the linear SVM divides the hyperplane by $\omega^T x + b = 0$ to distinguish the labeled data set, where $\omega$ is the normal vector and $b$ is the displacement term.

The distance between any point $x$ in the sample space and the hyperplane can be written as

$$r = \frac{\left|\omega^T (x + b)\right|}{\|\omega\|}. \tag{4}$$

| up_v | up_s | up_o | up_dif_v | up_dif_s | up_dif_o | down_v | down_s | down_o | own_dif_ | own_dif_ | own_dif_ | abs_dif_ | abs_dif_ | abs_dif_ | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| −0. 35797 | 0. 626395 | −0. 26101 | −0. 03055 | −1. 25263 | −0. 61971 | −0. 26454 | 0. 308747 | −0. 08715 | −0. 64797 | 0. 959104 | −0. 23376 | −0. 96101 | −0. 30024 | −0. 46374 | 1 |
| 1. 618194 | −1. 42883 | 0. 932333 | −0. 56345 | 1. 293378 | −0. 58119 | 0. 325458 | −2. 11207 | 2. 777052 | −0. 68292 | −0. 01154 | −0. 41908 | 1. 139127 | 0. 695528 | 1. 75901 | 1 |
| 0. 587152 | 0. 340272 | 0. 094922 | 1. 20274 | −0. 70182 | −1. 02375 | −0. 66166 | 0. 690003 | −0. 42753 | −0. 26358 | −1. 26354 | −0. 71554 | 0. 995936 | −0. 79279 | −0. 05608 | 1 |
| 1. 306733 | 0. 896161 | −0. 01299 | −1. 38564 | −0. 17417 | −0. 46032 | 1. 233148 | −0. 06932 | 0. 025452 | −0. 90423 | 0. 210748 | −0. 14847 | −0. 59508 | 0. 438862 | −0. 54873 | 1 |
| 0. 372352 | −2. 06269 | 1. 594566 | −1. 35519 | −0. 83324 | 1. 654453 | −1. 47858 | −1. 15658 | −0. 88232 | −0. 08886 | 1. 086471 | −0. 58745 | 1. 823264 | −0. 72722 | 1. 744638 | 1 |
| 0. 168291 | 0. 506585 | −0. 46551 | 0. 471904 | −0. 81691 | 0. 041384 | −1. 37646 | −1. 92292 | −0. 0363 | 0. 470254 | −0. 78588 | 0. 372971 | 1. 37778 | 2. 371049 | −0. 23084 | 1 |
| 1. 231553 | 0. 548245 | −0. 13686 | −0. 92887 | −1. 11472 | −0. 21109 | −0. 66166 | 0. 143606 | −0. 58176 | 1. 099256 | −0. 16166 | −0. 11704 | 1. 950545 | −0. 17101 | −0. 07998 | 1 |
| 1. 231553 | 0. 52015 | −0. 21716 | 0. 319646 | 0. 067754 | −0. 10036 | −0. 27589 | −0. 07157 | −0. 42592 | 0. 225642 | −0. 09079 | −0. 37185 | 1. 4096 | 0. 070218 | −0. 3254 | 1 |
| 1. 961874 | 1. 165155 | 0. 124981 | −1. 17248 | −0. 77047 | 0. 254079 | 1. 845839 | 0. 527096 | −0. 11114 | −0. 95083 | −0. 72805 | −0. 26248 | −0. 48371 | −0. 04094 | −0. 38884 | 1 |
| 1. 360433 | −2. 36298 | 2. 333208 | −0. 83751 | 0. 330979 | 4. 174152 | −0. 95666 | −3. 13368 | 2. 98878 | −0. 50819 | −1. 10678 | 3. 391442 | 2. 555131 | 1. 049727 | 0. 798217 | 1 |
| −0. 42241 | 0. 440146 | −0. 43721 | −0. 85274 | −1. 46675 | −0. 52668 | 0. 155266 | 0. 661721 | −0. 45338 | −0. 32182 | −0. 24746 | −0. 33702 | −1. 1042 | −0. 92546 | −0. 48289 | 1 |
| −0. 75535 | 0. 285662 | −0. 2705 | −1. 35519 | 0. 083988 | −0. 55676 | −0. 55954 | 0. 516482 | −0. 30366 | 2. 322317 | −0. 01227 | 0. 911983 | −1. 13602 | −0. 89652 | −0. 50965 | 1 |
| 1. 489313 | −0. 15358 | 0. 365465 | −0. 19803 | 0. 164962 | 0. 497135 | −0. 09435 | −1. 78248 | −0. 25894 | 1. 239035 | −0. 4738 | 0. 098564 | 1. 536881 | 1. 543322 | −0. 01524 | 1 |
| 1. 467833 | 0. 834087 | −0. 18245 | −1. 14203 | −0. 8546 | −0. 27352 | 0. 166612 | 0. 297302 | −0. 4287 | 0. 679922 | −0. 41383 | −0. 46005 | 1. 139127 | −0. 08075 | −0. 29249 | 1 |
| −0. 26131 | −3. 87334 | 5. 34829 | −0. 56345 | −0. 68677 | 4. 256876 | −1. 45588 | −1. 24396 | −0. 92407 | 0. 400365 | −0. 87019 | −0. 60705 | 0. 852745 | 0. 952264 | 5. 009321 | 1 |
| 0. 887872 | 1. 315064 | −0. 46278 | −1. 20293 | −0. 90746 | −1. 05294 | 1. 641609 | 1. 068389 | −0. 22264 | −0. 74116 | −0. 76605 | −0. 45744 | −0. 96101 | −0. 56948 | −0. 44433 | 1 |
| −2. 18377 | 0. 671476 | −1. 1686 | 0. 334872 | 0. 239919 | −0. 87324 | −2. 51108 | 0. 182772 | −1. 01382 | −0. 99742 | −1. 0001 | −0. 64233 | −0. 51553 | −0. 09823 | −0. 47471 | 1 |
| 1. 360433 | 0. 062559 | 0. 15608 | −1. 50744 | 0. 446779 | −0. 45341 | −0. 24185 | −0. 31503 | −0. 28399 | 0. 749811 | 0. 315881 | −0. 10724 | 1. 552791 | −0. 0775 | −0. 16632 | 1 |
| 1. 489313 | 0. 746277 | −0. 08093 | −0. 47209 | 0. 268719 | −0. 13234 | −0. 36666 | −1. 3948 | −0. 48993 | 0. 889589 | −1. 40465 | −0. 29345 | 1. 918725 | 1. 947669 | −0. 1361 | 1 |
| −0. 63721 | −0. 04861 | −0. 16601 | −1. 44654 | 2. 508246 | −0. 60973 | −0. 083 | 0. 441512 | −0. 4357 | 0. 097512 | 0. 105883 | −0. 18296 | −1. 12011 | −0. 71924 | −0. 27048 | 1 |
| 1. 317473 | 0. 375682 | 0. 05136 | −0. 95932 | 0. 768534 | −0. 05303 | −0. 59358 | −1. 33532 | −0. 64299 | 0. 784755 | −0. 10643 | −0. 57765 | 1. 982365 | 1. 50722 | 0. 150726 | 1 |

Figure 5: Low precision data set sample.

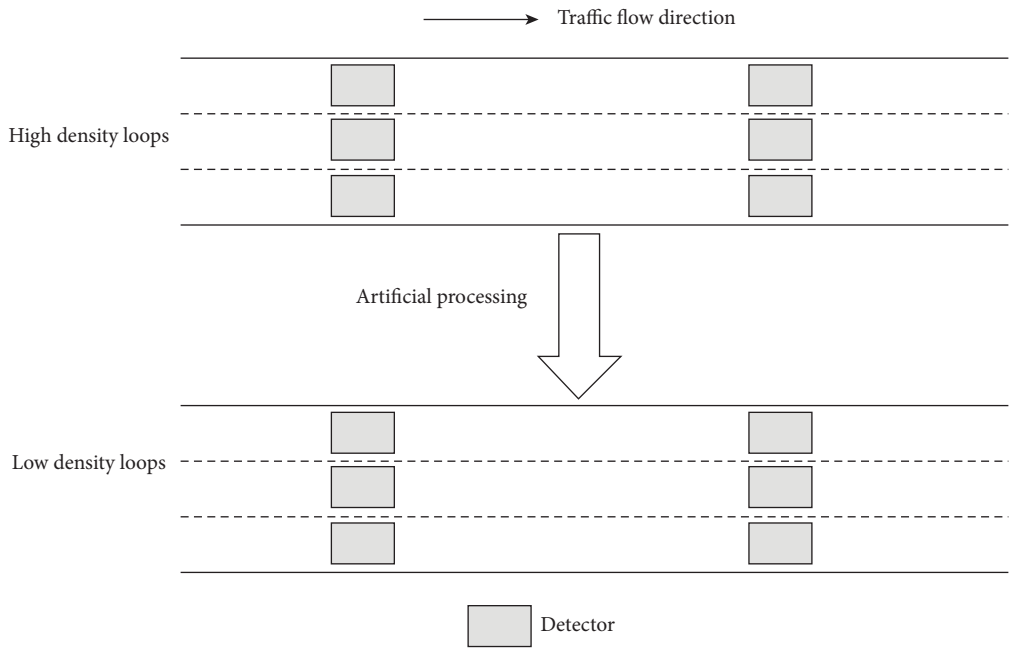


Figure 6: Schematic diagram of freeways with different loop densities.

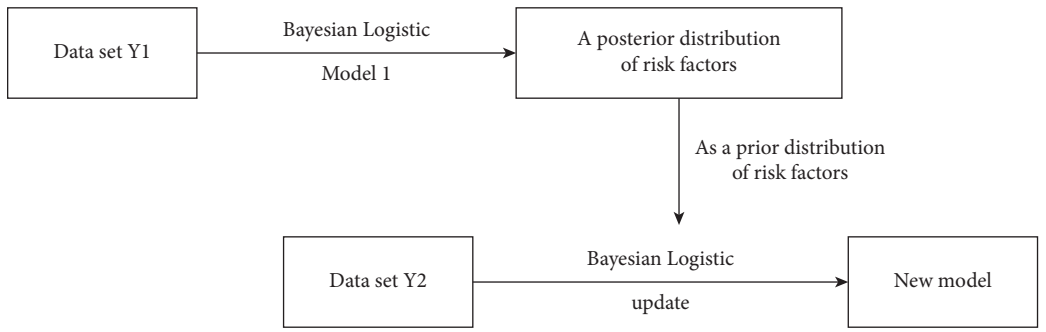


Figure 7: Schematic diagram of Bayesian updating method.

For labeled sample data sets $Y = (y_1, y_2, \ldots, y_n,)$, +1 is accident data, and −1 is nonaccident data. If it can be correctly classified, it can be

$$\begin{cases} \omega^T x_i + b \geq +1, y_i = +1, \\ \omega^T x_i + b \leq -1, y_i = -1. \end{cases} \tag{5}$$

When the sample dimension is high, it may lead to linear inseparability of sample data. The processing method of SVM for this situation is to raise the dimension of the sample data, convert the linear nonfraction data in the low dimensional space into linearly separable data in the high-dimensional space, and then use the linear SVM to find the optimal classification surface in the high-dimensional space.

*3.4. The Evaluation Index.* In the data classification model, accuracy can intuitively display the overall classification performance of the model, which is expressed as the proportion of the correctly classified sample results in the total sample among all samples as shown in the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

where TP represents the number of samples predicted to be positive; TN represents the number of samples that predicted negative classes as negative classes; FP represents the number of samples that predicted the negative category as the positive category; FN represents the number of samples that predicted positive classes as negative classes.

The confounding matrix shown in Table 2 can directly display the classification results of the model and calculate the corresponding true positive rate (TPR) and false-positive rate (FPR) indexes.

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \tag{7}$$

The receiver operating characteristic (ROC) curve can be drawn by using TPR and FPR. The ROC represents the curve of the prediction accuracy of the data set under different probability thresholds. The AUC value of the area under the ROC curve can be calculated to measure the quality of the model. The closer the AUC value is to 1, the better the performance of the model.

## 4. Results and Discussion

*4.1. Analysis of the Results from Small Sample Data Sets.* In order to study the influence of different sample size data sets on the established crash risk model, Bayesian Logistic regression is used to establish models for the extracted 5%, 10%, 20%, 30%, and 50% high precision data sets, respectively. Table 3 shows the significant risk factors screened by Logistic stepwise regression for each data set.

By comparing the significant risk factors of the models with different sample size data sets, it is found that the sample size does affect the real-time crash risk model. Different models not only share the same (i.e., the same impact factors) but also have their own characteristics. It provides a basis for subsequent analysis.

As can be seen from Table 3, the speed of the upstream loop (*up_s*) is a significant variable of each data set. It shows that, for different data sets, upstream is a significant factor affecting the occurrence of crashes, which plays an important role in explaining the causes of crashes. At the same time, there were differences in other risk factors among each data set. Some risk factors were significant in one small sample, but not in others. This shows that each small sample data set has different characteristics and has certain differences for the establishment of the real-time crash risk model.

Figure 8 is the ROC curve and AUC value diagram of the real-time crash risk model established by the Bayesian Logistic regression method with different small sample data sets.

As can be seen from Figure 8, the change of the AUC value of the real-time crash risk model established by Bayesian Logistic regression does not increase with the increasing of sample size of the data set but is in a state of fluctuation. However, the overall trend of AUC value is decreasing.

For each collision, the traffic flow state is different. When the sample size of the data set is different, the structure of the data set is more complex. The significant risk factors screened by the Bayesian Logistic model established for each set of data mainly explain the predictive classification effect of the data set. And the significant risk factor of each data set is the optimal combination screened by the model. Therefore, there will be different results of collision precursors under different data sets. As mentioned above, as the sample size increases, the data set structure becomes more complex. Under the given combination of significant risk factors in different data sets, the probability of accidents is more complex, so the AUC index of the model will be reduced.

With the increase of sample size, the number of traffic crashes also increases. At this point, the diversity of traffic flow states of traffic crashes increases. From the screening of risk factors, it can be seen that the combination of risk factors changes with different sample sizes. Both the common significance factors (such as *up_s*) and the unique significance factors of each data set are included. Different traffic flow states make the data structure diverse. Therefore, the accuracy of models based on different data sets may be reduced. At the same time, due to the increase of data volume, the amount of traffic crash data increases. Although the AUC value of the overall model decreases, it is still around 0.7 or even a little higher, indicating that the number of traffic crashes correctly classified increases, too. It also indicates that while the sample size increases, though the sample structure is diverse, the law of data can be extracted as the sample size increases. This is the improved classification performance of the model. The increase of sample size can improve the prediction performance of the real-time crash risk model.

Table 2: Confusion matrix.

| | | Predict | | Total |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual | Positive | TP | FN | TP + FN |
| | Negative | FP | TN | FP + TN |
| Total | | TP + FP | FN + TN | TP + FN + FP + TN |

Table 3: Significant variables in small sample data sets.

| Data sets | Risk factors |
|---|---|
| 5% small sample | *up_s* |
| 10% small sample | *up_s*, *down_v*, *abs_dif_s*, and *up_dif_v* |
| 20% small sample | *up_s*, *down_s*, *up_o*, and *down_v* |
| 30% small sample | *up_s* and *down_dif_v* |
| 50% small sample | *up_s*, *down_dif_v*, *up_dif_s*, and *abs_dif_o* |

At the level of significance of 0.05.

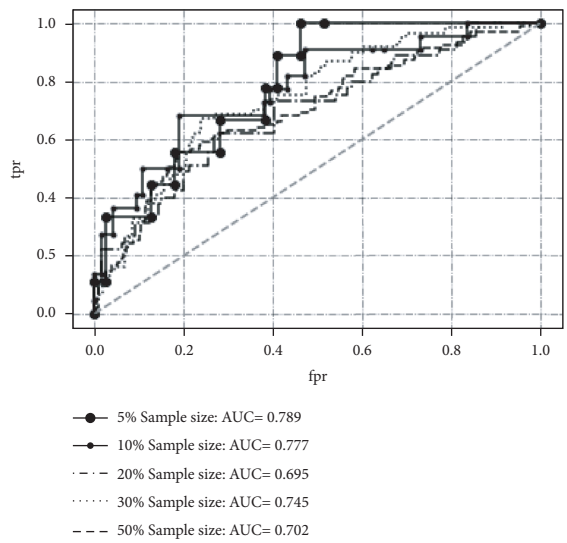


Figure 8: The ROC curve and AUC value of the model were established with a small sample data set.

### 4.2. Reliability Verification of Model Transferability

*4.2.1. Low Precision Data Set and High Precision Data Set Risk Model Comparison.* Stepwise Logistic regression is used to screen the factors that had a significant impact on the risk model, and the Bayesian method is used to estimate the coefficients of the model. Table 4 shows the model comparison after the coefficient estimation of significant risk factors.

As can be seen from Table 4, there are partially the same explanatory variables with significant levels between low precision and high precision data sets, such as *abs_dif_o* and *up_s*. These two explanatory variables indicate that, in the two data sets, both the speed of upstream loop and the absolute value of occupancy difference between upstream and downstream loop can effectively explain the causes of crashes. The difference between the two sets of data sets is that *up_dif_o* in the low precision data set has a significant explanatory effect on the model, while *down_dif_v* and *up_dif_s* in the high precision data set have a stronger explanatory effect on the model. In the coefficient estimation, it can be found that the estimated 95% confidence interval of each coefficient does not contain 0, indicating that the coefficient estimation is significant. The average upstream speed (*up_s*) coefficient of the two models is negative, indicating that, within the specified driving speed range of expressways, the decrease of average upstream speed by one unit will lead to an increase in crash risk.

Each piece of data is classified, and the accuracy of the model established by the two sets of data sets is shown in Table 5.

At the same time, the confusion matrix of two data sets for model classification and prediction is shown in Tables 6 and 7.

The ROC curve and AUC values of the model established on the basis of the two data sets are shown in Figure 9.

Through the comparison of the above indicators, it can be found that when the data set established with relatively sparse loop density is used to establish the Bayesian Logistic regression model, the classification accuracy of the model is 70.68%, slightly lower than the classification accuracy of the high precision data set with large loop density which is 73.30%. However, the model AUC value of low precision data set is 0.656, which is much smaller than that of high precision data set. The reasons are as follows: there are few explanatory factors in the low precision data set, and the data information is lost, which affects the accuracy of the model to some extent. In contrast, when the loop density is larger, more traffic flow information can be collected, and traffic variables that have a significant impact on traffic crashes can be screened out, thus making the model more accurate.

*4.2.2. The Application of the Model of the High Precision Data Set-Based Model Low Precision Data Set.* Applying the model of high precision data set to low precision data set, the classification accuracy is 69.3%, and the confusion matrix is shown in Table 8.

The ROC curve and AUC value obtained are shown in Figure 10.

Directly applying the model established by high precision data sets to low precision data sets, the classification results are worse than those established by previous low precision data sets. When using Logistic regression to screen variables, the optimal variable combinations of the two data sets are different. In the process of parameter estimation, the model obtained is the best fit of the best variables under each set of data. Therefore, when applied to other data sets, there

Table 4: Significant risk factors and parameter estimates for the two data sets models.

| Variable | Low precision data set | | High precision data set | |
|---|---|---|---|---|
| | Mean | Confidence interval | Mean | Confidence interval |
| Constant | −1.477 | (−1.657, −1.294) | −1.5734 | (−1.755, −1.391) |
| *abs_dif_o* | 0.183 | (0.008, 0.358) | 0.276 | (0.102, 0.456) |
| *up_s* | −0.299 | (−0.507, −0.095) | −0.608 | (−0.784, −0.427) |
| *down_dif_v* | — | — | −0.209 | (−0.387, −0.037) |
| *up_dif_s* | — | — | 0.250 | (0.087, 0.405) |
| *up_dif_o* | 0.257 | (0.060, 0.454) | — | — |

Table 5: Classification accuracy of Bayesian Logistic regression.

| Data sets | Low precision data set (%) | High precision data set (%) |
|---|---|---|
| Accuracy | 70.68 | 73.30 |

Table 6: Low precision data set confusion matrix.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 74 | 87 |
| Negative (actual) | 149 | 495 |

Table 7: High precision data set confusion matrix.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 116 | 75 |
| Negative (actual) | 180 | 584 |

will be inapplicable situations. It can be seen that direct transplantation of the model cannot achieve a better prediction classification effect.

*4.2.3. Bayesian Updating towards High Precision Data Set Model and Low Precision Data Set Model.* (a) The Bayesian updating method was used to update and transplant the model established by the original high precision data set. The posterior distribution of parameter estimation of the low precision data set model variables was regarded as the prior distribution of parameter estimation of the high precision data set model and then updated it. The results obtained are shown in Table 9.

The classification accuracy of the updated high precision data model is 68.94%, and the confusion matrix is shown in Table 10. ROC curve and AUC value obtained are shown in Figure 11.

(b) The Bayesian updating method was used to update and transplant the model established by the original low precision data set. The posterior distribution of parameter estimation of the high precision data set model variables was regarded as the prior distribution of parameter estimation of the low precision data set model and then updated it. The results obtained are shown in Table 11.

The classification accuracy of the updated low precision data set model is 70.83%, and the confusion matrix is shown in Table 12. ROC curve and AUC value obtained are shown in Figure 12.

By updating the model established by the high precision data set, it can be found that the prediction accuracy of the model cannot be effectively improved when the model is applied to the low precision data set before and after updating. The prediction accuracy is 69.3% before the update and 68.94% after the update, which decreases by 0.36%, and the AUC value decreases by 0.002.

By updating the model established by the low precision data set, it can be found that the prediction accuracy of the model is 70.68% and 70.83%, respectively, before and after the model is applied to the low precision data set, and the classification accuracy is improved by 0.15%. In addition, the AUC value increases from 0.656 to 0.657.

The classification accuracy of the model is 70.68% in the low precision data set. Based on the evaluation index of the model, the classification accuracy of the model transplant results is improved to a certain extent. Therefore, 69.3% could not meet the requirements, while the result of another model transplantation reached the requirements of 70.83%. In comparison, the improvement of the model is smaller, only 0.15%. However, in the field of traffic safety, it will have practical application significance to improve certain accuracy. In the follow-up research, better models or methods can be further proposed to make the results of model transplantation better. It can be seen from the above that the Bayesian updating method can improve the model transplantation effect to a certain extent, but the overall effect is limited, indicating that this method can indeed carry out model transplantation. The reason for the limited improvement may be that the most significant factor with explanatory effect has been screened out during the stepwise regression, and the difference between the parameters estimated by the Bayesian method and the parameters estimated by maximum likelihood estimation is small. At this time, the parameters of the model have become an excellent combination of parameter values. In the process of Bayesian update, the prior information of the model has little influence on it, so the overall improvement effect of the model is small.

In addition, it is necessary to determine the updated model object. Through the above study and comparison, it can be concluded that updating models with low precision data sets will obtain models with higher prediction performance.

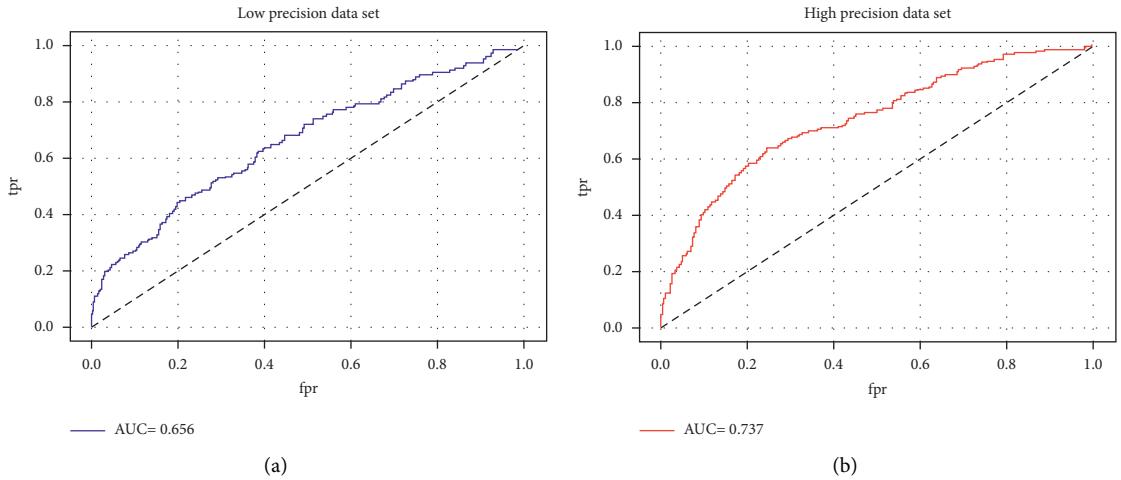


Figure 9: ROC curve and AUC values of the model were established for the two sets of data.

Table 8: The high precision data set model is applied to the low precision data set.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 79 | 82 |
| Negative (actual) | 165 | 479 |

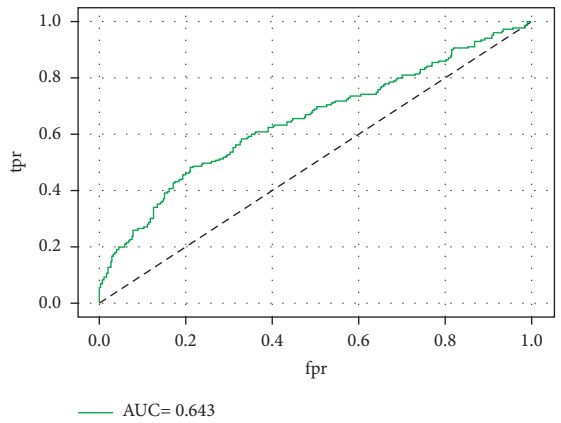


Figure 10: The high precision data set model is applied to the low precision data set.

Table 9: Results of the updated high precision data set model.

| Variable | Before | | After | |
|---|---|---|---|---|
| | Mean | Confidence interval | Mean | Confidence interval |
| Constant | −1.573 | (−1.391, 2.474) | −1.540 | (−1.715, −1.365) |
| *abs_dif_o* | 0.276 | (0.102, 0.456) | 0.267 | (0.148, 0.385) |
| *up_s* | −0.608 | (−0.784, −0.427) | −0.0.494 | (−0.624, −0.364) |
| *down_dif_v* | −0.209 | (−0.387, −0.037) | −0.212 | (−0.387, −0.038) |
| *up_dif_s* | 0.250 | (0.087, 0.405) | 0.236 | (−1.715, −1.365) |

*4.3. Classification Prediction Model Based on SVM.* In order to be consistent with previous research methods, data sets were not divided into training data sets and test data sets. Classification prediction models for high precision and low precision data sets are established based on SVM, and the classification accuracy and confusion matrix are obtained as in Tables 13–15.

ROC curve and AUC values are shown in Figure 13.

Table 10: Confusion matrix of the updated high precision data set model.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 81 | 80 |
| Negative (actual) | 170 | 474 |

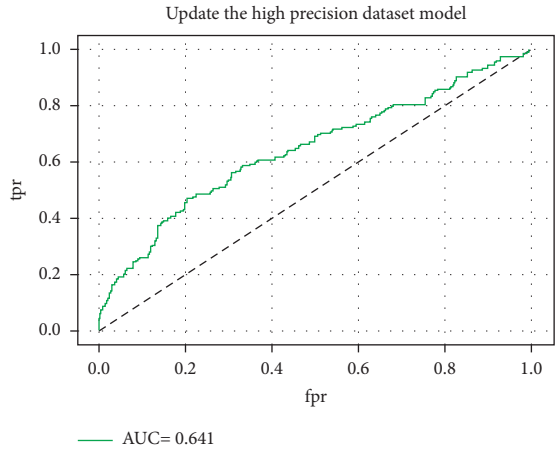


Figure 11: ROC curve and AUC value of the updated high precision data set model.

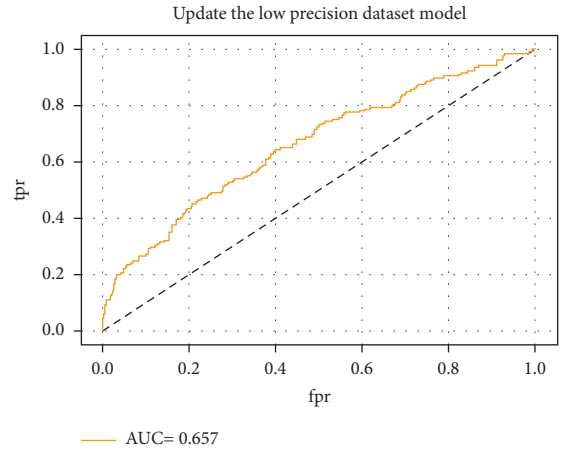


Figure 12: ROC curve and AUC value of the updated low precision data set model.

Table 11: The results of the updated low precision data set model.

| Variable | Before | | After | |
|---|---|---|---|---|
| | Mean | Confidence interval | Mean | Confidence interval |
| Constant | −1.477 | (−1.650, −1.283) | −1.470 | (−1.663, −1.290) |
| *abs_dif_o* | 0.183 | (0.012, 0.354) | 0.189 | (0.023, 0.352) |
| *up_s* | −0.300 | (−0.503, −0.098) | −0.307 | (−0.512, −0.110) |
| *up_dif_o* | 0.252 | (0.053, 0.451) | 0.253 | (0.052, 0.446) |

Table 12: Confusion matrix of the updated low precision data set model.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 75 | 86 |
| Negative (actual) | 151 | 493 |

Table 13: SVM classification accuracy.

| | Low precision data set (%) | High precision data set (%) |
|---|---|---|
| Accuracy | 76.9 | 78.7 |

Table 14: Confusion matrix of the low precision data set.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 81 | 80 |
| Negative (actual) | 46 | 598 |

Table 15: Confusion matrix of the high precision data set.

| | Positive (predict) | Negative (predict) |
|---|---|---|
| Positive (actual) | 94 | 97 |
| Negative (actual) | 57 | 707 |

By analyzing these two groups of data sets with SVM model, this paper finds that when the loop density is relatively sparse, the precision of the model does have some influence, low accuracy of the data set to establish the SVM model accuracy is 76.9%, the AUC value is 0.8, high precision data set to establish accuracy of SVM is 78.7%, and the AUC value is 0.82. The comparison between the two models shows that the high precision data set is better for modeling. Compared to other machine learning models, Shen considered the weather variables when establishing the random forest real-time accident risk model [33]. In this model, the accuracy of the model reached 82.1%. In this study, the authors screened the characteristics of the data and took the weather into account. Compared with the support vector machine model, the accuracy was improved by 3.4%. In further researches, the data could be processed accordingly, and the parameters of the support vector machine could be adjusted to achieve higher prediction performance.

Compared with the Bayesian Logistic regression model, in the case of the same low precision data set, the overall prediction performance of the established SVM model is better, with the classification accuracy improved by 6.22%, and the AUC value is improved by 0.144. When loop density is small and the loop data information is not rich, a real-time crash risk model based on Bayesian Logistic regression can be established, which can effectively filter out the significant risk factors for crashes and can be explained in detail and quantify the corresponding risk factors. However, strict and

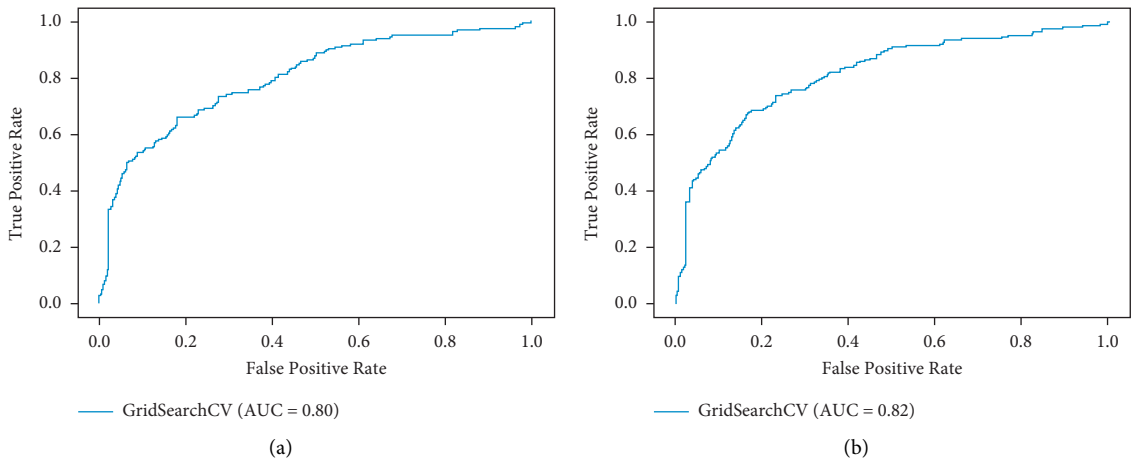


Figure 13: ROC curve and AUC value of low precision and high precision data sets.

mathematical relations limit the overall prediction effect of the model. SVM is a black box machine learning algorithm, which can effectively learn the effects of features on the results and reflect them into the prediction results.

Therefore, when the data set is not accurate enough, it is recommended to use the machine learning algorithm to establish a model to classify and predict the crash risk. When the data accuracy is good, the statistical Logistic regression method can be used to screen out significant risk variables to explain the model and classify the crash risk prediction.

## 5. Conclusions

Considering the influence of limited data conditions on the real-time freeway traffic crash risk model, this paper constructed high precision data set, low precision data set, and small sample data set. These data sets were modeled and analyzed based on Bayesian Logistic regression, and the reliability of real-time crash risk model transplantation based on Bayesian update was verified. Finally, the advantages and disadvantages of the model established by Bayesian Logistic and SVM were compared. The main conclusions of this paper are as follows:

(1) The significant risk factors of Bayesian Logistic regression established under various sample sizes are different. With the increasing of sample size, the evaluation index of the model decreases. However, the overall performance of the model improves. The increase of sample size can effectively improve the classification and prediction performance of the model.

(2) When the loop detector density of the collected data is small, the prediction performance of the Bayesian Logistic regression model based on low precision data set is weaker than that of the Bayesian Logistic regression model based on high precision data set. In addition, significant risk factors are significantly different in the two models, indicating that Bayesian Logistic regression is not suitable for low precision data set.

(3) Based on the Bayesian updating method, the validity of model migration is verified. Applying the posterior distribution of significant variable parameters of the Bayesian Logistic model based on high precision data set to low precision data set, this approach can improve the prediction performance of the Bayesian Logistic model using low precision data set.

(4) Compared with Bayesian Logistic regression, the crash risk model based on SVM has higher prediction performance. Even under the condition of low precision data set, its prediction performance is significantly improved compared with that of Bayesian Logistic regression, indicating that SVM is a better choice under the condition of insufficient data precision. However, SVM cannot effectively interpret the cause of crash risk. When the data quality is high, Bayesian Logistic regression can be used for modeling and prediction, and the crash risk can be well explained.

In this paper, Bayesian Logistic regression and support vector machine are applied to analyze the impact of various data sets on the traffic crash risk model. Further, other machine learning methods and the enhancement effect of feature engineering on the establishment for the crash risk model can be studied. Some new methods of crash risk model transplantation should also be studied in the future.

## Data Availability

The data used to support the findings in this study are available from the corresponding authors upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

## Acknowledgments

This research was supported by the China Postdoctoral Science Foundation (2021M700333).

## References


[1] W. Chu, S. Su, and L. Zhou, "Analysis and suggestions on traffic safety hazards of pure electric vehicles," *Road Traffic Management*, vol. 35, no. 4, pp. 32-33, 2020.

[2] W. Li, Z. Pu, Y. Li, and X. Ban, "Characterization of ride-splitting based on observed data: a case study of Chengdu, China," *Transportation Research Part C: Emerging Technologies*, vol. 100, pp. 330–353, 2019.

[3] Y. Yang, *Research on the Method of Freeway Crash Risk Identification and Comprehensive Traffic Safety Evaluation Considering the Regional Type difference*, Beijing Jiaotong University, Beijing, China, 2020.

[4] Z. Yuan, C. Lou, and Y. Yang, "Analysis of highway traffic accidents causes under time differences," *Journal of Beijing Jiaotong University*, vol. 45, no. 3, pp. 1–7, 2021.

[5] B. Tian, C. Liang, and Y. Bao, "Spatial and temporal distribution characteristics of traffic accidents on mountain highways and safety improvement measures," *Journal of Wuhan University of Technology*, vol. 42, no. 6, pp. 1014–1018, 2018.

[6] W. Y. Mergia, D. Eustace, D. Chimba, and M. Qumsiyeh, "Exploring factors contributing to injury severity at freeway merging and diverging locations in Ohio," *Accident Analysis and Prevention*, vol. 55, pp. 202–210, 2013.

[7] X. Ye, R. M. Pendyala, V. Shankar, and K. C. Konduri, "A simultaneous equations model of crash frequency by severity level for freeway sections," *Accident Analysis and Prevention*, vol. 57, pp. 140–149, 2013.

[8] N. Haghighi, X. C. Liu, G. Zhang, and R. J. Porter, "Impact of roadway geometric features on crash severity on rural two-lane highways," *Accident Analysis and Prevention*, vol. 111, pp. 34–42, 2018.

[9] M. Osman, S. Mishra, and R. Paleti, "Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: accounting for unobserved heterogeneity and age group differences," *Accident Analysis and Prevention*, vol. 118, pp. 289–300, 2018.

[10] P. Xu, H. Huang, N. Dong, and S. C. Wong, "Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach," *Accident Analysis and Prevention*, vol. 98, pp. 330–337, 2017.

[11] W. Wang, Z. Yuan, Y. Yang, X. Yang, and Y. Liu, "Factors influencing traffic accident frequencies on urban roads: a spatial panel time-fixed effects error model," *PLoS One*, vol. 14, no. 4, Article ID 0214539, 2019.

[12] Y. Yang, "Analysis of the factors influencing highway crash risk in different regional types based on improved Apriori algorithm," *Advances in Transportation Studies*, vol. 17, pp. 165–178, 2019.

[13] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Analysis and Prevention*, vol. 45, pp. 478–486, 2012.

[14] Y. Guo, Z. Li, and T. Sayed, "Analysis of crash rates at freeway diverge areas using Bayesian to bit modeling framework," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 4, pp. 652–662, 2019.

[15] X. Lv and Q. Wang, "Analysis of traffic accident severity on circular expressway based on logistic," *Journal of Wuhan University of Technology (Transportation Science and Engineering)*, vol. 63, pp. 1–8, 2021.

[16] C. H. Panagiotis, "Anastasopoulos. random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis," *Analytic Methods in Accident Research*, vol. 11, pp. 17–32, 2016.

[17] Q. Hou, X. Huo, J. Leng, and Y. Cheng, "Examination of driver injury severity in freeway single-vehicle crashes using a mixed logit model with heterogeneity-in-means," *Physica A: Statistical Mechanics and Its Applications*, vol. 531, Article ID 121760, 2019.

[18] Y. E. Fan and D. Lord, "Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models," *Analytic Methods in Accident Research*, vol. 1, pp. 72–85, 2014.

[19] Y. Yang, Z. Yuan, J. Chen, and M. Guo, "Assessment of osculating value method based on entropy weight to transportation energy conservation and emission reduction," *Environmental engineering and management journal*, vol. 16, no. 10, pp. 2413–2423, 2017.

[20] X. Chen, J. Ling, S. Wang, and Y. S. Yang, "Ship detection from coastal surveillance videos via an ensemble Canny-Gaussian-morphology framework," *Journal of Navigation*, vol. 74, no. 6, pp. 1252–1266, 2021.

[21] Z. Pu, Z. Li, Y. Jiang, and Y. H. Wang, "Full Bayesian before-after analysis of safety effects of variable speed limit system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1–13, 2020.

[22] H. Yang, C. Liu, M. Zhu, X. Ban, and Y. Wang, "How fast you will drive? predicting speed of customized paths by deep neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–11, 2021.

[23] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1784, no. 1, pp. 1–8, 2002.

[24] M. Abdel-Aty, N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia, "Predicting freeway crashes from loop detector data by matched case-control logistic regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1897, no. 1, pp. 88–95, 2004.

[25] H. Yang, R. Ke, Z. Cui, Y. Wang, and K. Murthy, "Toward a real-time smart parking data management and prediction (SPDMP) system by attributes representation learning," *International Journal of Intelligent Systems*, vol. 36, pp. 1–34, 2021.

[26] G. Fountas and P. C. Anastasopoulos, "A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities," *Analytic Methods in Accident Research*, vol. 15, pp. 1–16, 2017.

[27] J. Sun and J. Sun, "Proactive assessment of real-time traffic flow accident risk on urban expressway," *Journal of Tongji University*, vol. 42, no. 6, pp. 873–879, 2014.

[28] J. You, J. Wang, T. Tang, and S. Fang, "Support vector machines approach for predicting real-time rear-end crash risk on freeways," *Journal of Tongji University*, vol. 45, no. 3, pp. 355–361, 2017.

[29] C. Xu, P. Liu, W. Wang, and Z. B. LI, "Real time crash risk prediction model on freeways under nasty weather conditions," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 43, no. 1, pp. 68–73, 2013.

[30] X. Chen, H. Chen, Y. Yang et al., "Traffic flow prediction by an ensemble framework with data denoising and deep learning model," *Physica A: Statistical Mechanics and Its Applications*, vol. 565, Article ID 125574, 2021.

[31] X. Ma, B. Fan, and S. Chen, "Evaluation and analysis model for freeways crash risk based on real-time traffic flow," *Journal*

*of South China University of Technology*, vol. 49, no. 8, pp. 19–25+34, 2021.

[32] C. Xu, W. Wang, P. Liu, R. Guo, and Z. Li, "Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models," *Transportation Research Part C: Emerging Technologies*, vol. 38, pp. 167–176, 2014.

[33] J. Shen, *Real-time Risk Prediction and Spatiotemporal Impact Analysis for Freeway Accident*, Southeast University, Dhaka, Bangladesh, 2017.