A Machine Learning-Based Framework for Structured Traffic Risk Prediction Using Spatiotemporal Data

Guangxiao Li, Tao Wang

Information and Archives Technology Research Office, Jinan Special Structure Research Institute, Aviation Industry Corporation of China, Jinan 250100, Shandong, China

Abstract: ***The timeliness and effectiveness of traffic accident rescue operations are of critical importance for ensuring the safety of citizens' lives and property. Accurate prediction of traffic accident risks, coupled with prompt on-site investigation and rescue upon identification of potential hazards, would significantly enhance response efficiency and effectively safeguard public safety. Existing traffic risk prediction models face two principal limitations. First, their accuracy is often compromised by analysts' insufficient consideration of factors that induce traffic accidents, as well as constraints imposed by experiential or domain knowledge boundaries. Second, most models assess risk solely through the probability of traffic accident occurrence, while neglecting the impact of accident severity on overall risk levels. To address these limitations, this study proposes a structured real-time traffic risk prediction model that decomposes overall traffic risk into two hierarchical components: accident occurrence probability and accident severity. This dual-structure approach facilitates a more nuanced understanding of how different factors contribute to overall risk and offers distinct advantages in adapting to the dynamic nature of transportation environments. The methodological framework proceeds as follows. First, necessary data preprocessing is performed on the target data objects. Subsequently, data mining techniques and in-depth analysis are applied to examine various spatiotemporal factors influencing traffic accidents, utilizing analytical methods including autocorrelation analysis and Pearson correlation analysis. Feature selection is then conducted based on the analytical results. Finally, the traffic risk prediction model is constructed using the PU-Learning algorithm in combination with a random forest model. Experimental analysis demonstrates that the proposed model achieves superior predictive performance compared to benchmark models, confirming its effectiveness in addressing the challenges inherent in dynamic transportation environments.***

Keywords: Data mining; Traffic risk prediction; Machine learning.

1. INTRODUCTION

With the general improvement of people's living standards, the demand for automobiles has increased year by year, and the number of traffic accidents has also continued to rise. According to data from the World Health Organization, more than 1.3 million people worldwide die in road traffic accidents each year, with one person killed every 24 seconds. It is the leading cause of death for young people and children, and 20 to 50 million people suffer serious physical injuries, causing huge economic losses to individuals, families, and the entire country. How to effectively predict and respond to traffic accidents through emerging technologies, eliminate potential traffic safety hazards at an early stage, improve the ability to prevent road traffic accidents, and reduce the occurrence of such tragedies has become one of the hot research topics for governments and scientific research institutions at home and abroad.

There are many factors causing road accidents, including spatial correlation, temporal dynamic interaction, and external influences. If historical data and data statistics can be used to explore the internal relationship between road traffic accidents and dynamic influencing factors, scientific estimation and prediction of potential traffic accident risks can be made. This allows for rapid response to accident risks by proactively taking corresponding traffic safety management measures before accidents occur, thereby reducing the probability or severity of accidents. Therefore, traffic accident risk prediction is essentially a Spatio-Temporal Data Mining (STDM) problem. Tu [1] introduced a platform-aware framework, AutoNetTest, for intelligent 5G network test automation and issue diagnosis. In the domain of green logistics, Meng et al. [2] applied deep learning to optimize site selection and path planning in warehousing systems. Wu [3] focused on fault detection and prediction models to optimize resource usage in cloud infrastructure, while Chen [4] emphasized the importance of efficient and scalable data pipelines for data processing in gig economy platforms. In the processing of medical texts within legal documents, Yuan [5] employed transformer architecture to improve efficiency. Li, Wang, and Zhang [6] explored named entity recognition for smart city data streams to enhance visualization and interaction. For

federated learning security, Deng and Yang [7] proposed multi-layer defense strategies and privacy-preserving enhancements against membership reasoning attacks. Lin et al. [8] developed a Bayesian framework for modeling multivariate degradation data with dynamic covariates. Yi [9] addressed real-time fair-exposure ad allocation for small and medium-sized businesses and underserved creators using contextual bandits-with-knapsacks. In photonics, Tang et al. [10] designed and optimized a shallow-angle grating coupler for vertical emission from indium phosphide devices. Deng [11] proposed a homomorphic encryption-based mechanism for data integrity verification and anti-tampering in cloud storage environments. Expanding to financial infrastructure, Mehta et al. [12] developed a national AI security framework for protecting financial systems. Zhou [13] applied gradient boosting trees to diagnose bottlenecks in international automotive sales funnels, with evidence from cross-regional team efficiency evaluation. Wensi [14] explored AI-enabled data visualization marketing for automated production lines to build customer trust and improve lead-to-order conversion. Li [15] presented AI-based methods for predicting automation equipment lifecycle costs as a pathway to enhancing customer lifetime value. In object detection, Ren [16] developed adaptive multi-scale fusion for infrared and visible object detection within the YOLOv8 framework. Ximeng and Yiming [17] employed offline conservative reinforcement learning to balance fraud risk and customer friction in transaction authorization. Yang and Zhang [18] introduced edge-enabled real-time fraud detection for network lending terminals under low-latency constraints. Wang et al. [19] investigated AI end-to-end autonomous driving systems. Lastly, Chen et al. [20] addressed one-stage object referring with gaze estimation in computer vision applications.

2. DATA ANALYSIS AND FEATURE SELECTION

2.1 Dataset Description

The data is selected from the US traffic accident dataset (version 5.0) released by Moosavi et al. on Kaggle, which contains nationwide traffic accident data from 49 states in the United States from February 2016 to June 2020, with approximately 1.5 million accident records. Each accident record includes 47 attributes describing traffic accidents. Since this study mainly focuses on urban-level traffic risk prediction, the city of Los Angeles was selected as the main research area after comprehensive consideration of multiple indicators such as city size, traffic accident data volume, population density, traffic travel demand, and urban influence. The main attributes of the dataset and their descriptions are shown in Table 1:

Table 1: Main attributes and descriptions

	Attribute			**Description**
Severity Level	Severity			Accident severity (value from 1 to 4)
Occurrence Time	Start Time	End Time		Accident start time and end time
Location Information	Start Lat	Start Lng	State	Accident GPS coordinates, state and city, street, nearest airport weather station
	City	Street	Airport Code	
Weather Conditions	Temperature	Humidity	Wind Direction	Temperature, humidity, wind direction, wind speed, atmospheric pressure, visibility, etc. at the accident location
	Wind Speed	Pressure	Visibility	
POI Markers	Junction	Traffic Signal	Crossing	Whether there are junctions, traffic signals, speed bumps, and other markers at the accident location
	Bump	Amenity	Railway	

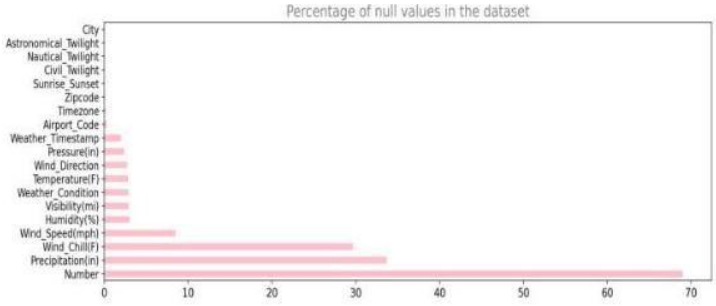


Figure 1: Missing Condition of the Dataset

2.2 Data Preprocessing

2.2.1 Missing Value Analysis and Handling

Figure 1 visualizes the data missing condition of the dataset. It can be seen that there are 19 columns with missing data, among which the three columns "Number", "Wind_Chill (F)", and "Precipitation (in)" have the most serious missing values, with 29.3%, 33.8%, and 69.1% missing respectively.

(1) Data Deletion

These three columns with a very high number of missing values have little correlation with traffic risk prediction and are directly deleted. In addition, compared with the overall sample, the missing values in feature attributes such as "City" and "Timezone" are very few, so data deletion is also directly used here for processing.

(2) Data Imputation

For weather feature columns with very few missing values, such as "Pressure (in)", "Wind_Direction", "Temperature (F)", "Visibility (mi)", "Humidity (%)", and "Wind_Speed (mph)", the data imputation method is used to handle missing items to retain the valuable information implied. The processing steps are as follows:

Select the location feature "Airport_Code" and the time feature "Start_Month", and group the weather features by location and time.

For discrete weather features like "Wind Direction", their missing values are replaced with the mode of each group, while for other continuous weather features, the median is used for filling.

2.2.2 Data Transformation

(I) Weather Data Simplification

To make it easier for the model to grasp and memorize the characteristics and variation rules of different weather types, similar weather types are merged and simplified into 7 items according to the inclusion relationship shown in Table 2:

Table 2: Simplified Weather Data

Weather Condition	**Values**
Clear	Clear
Cloud	Cloudy, Funnel Cloud, Scattered Clouds, Overcast
Rain	Light Rain, Rain, T-Storm, Light Thunderstorm, Light Thunderstorms, Thunderstorm, Thunderstorms,
Heavy Rain	Rain Shower, Rain Showers, Light Rain Shower, Light Rain Showers, Heavy Rain, Heavy Rain Showers, Heavy T-Storm
Snow	Snow, Light Snow, Light Snow Grains, Snow Grains, Low Drifting Snow, Ice Pellets, Light Ice Pellets, Sleet, Light Sleet,
Heavy Snow	Heavy Snow, Heavy Sleet, Heavy Ice Pellets, Light Snow Shower, Snow Showers, Squalls
Fog	Fog, Light Fog, Partial Fog, Patches of Fog, Shallow Fog

(2) Simplification of wind direction data

Similarly, simplify the wind direction data according to Table 3:

Table 3: Simplification of wind direction data

Wind Direction	**Values**
E	E, East, ESE, ENE
W	W, West, WSW, WNW
S	S, South, SSW, SSE

N	N, North, NNW, NNE
NE	NE
SW	SW
NW	NW
CALM	Calm
VAR	Var, Variable

2.2.3 Clustering-based Processing

Although POIs, streets, or population can reflect spatial changes to a certain extent, there are still significant differences in traffic accident patterns (causes, types, traffic enforcement) between different locations in the city (i.e., the central core area and residential areas). To address this, before modeling and prediction, spatial heterogeneity in the data was first processed using the K-Means clustering method, which divides the sample set into multiple categories based on internal data similarity.

The K-Means algorithm measures the similarity between samples by distance and then groups similar samples into the same cluster. Experiments showed that the optimal results were achieved when $k = 7$, so the number of clusters was set to 7 here. Figure 2 shows the clustering results for traffic accidents in Los Angeles in 2020, where seven differently colored regions represent categories with different degrees of similarity, and red dots indicate the centers of the seven clusters numbered 1-7. A new attribute "cluster_id" was then added based on the cluster center number each sample belongs to.

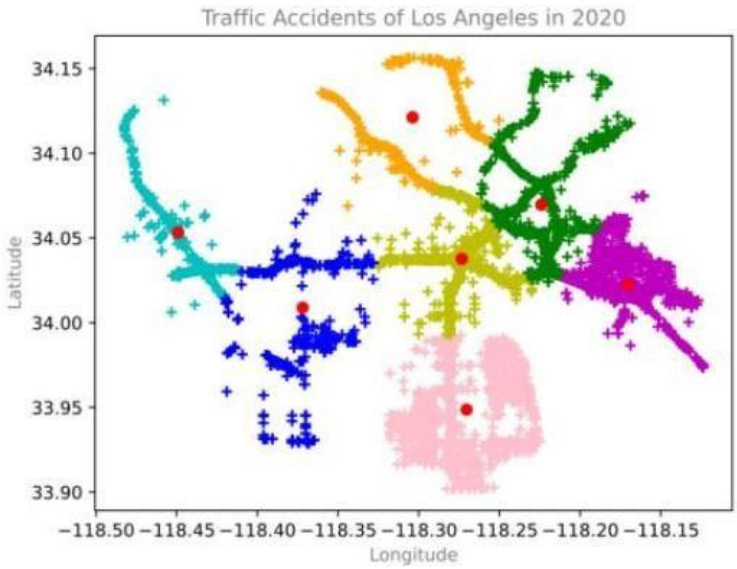


Figure 2: Division of Similarity Degree Based on K-Means

2.3 Data Analysis and Feature Selection

2.3.1 Analysis of Feature Importance

Figure 3 shows the top 22 important features calculated by the Extremely Randomized Trees (ERT) ensemble learning model. It can be seen that in the traffic accident severity model, the most important predictive features are Hour, Weekday, and Year. Other important predictors include Minute, Start_Lat, Temperature, Start_Lng, Pressure, Humidity, and Visibility, as well as the self-added road grade (street_flag) and cluster number (cluster_id).

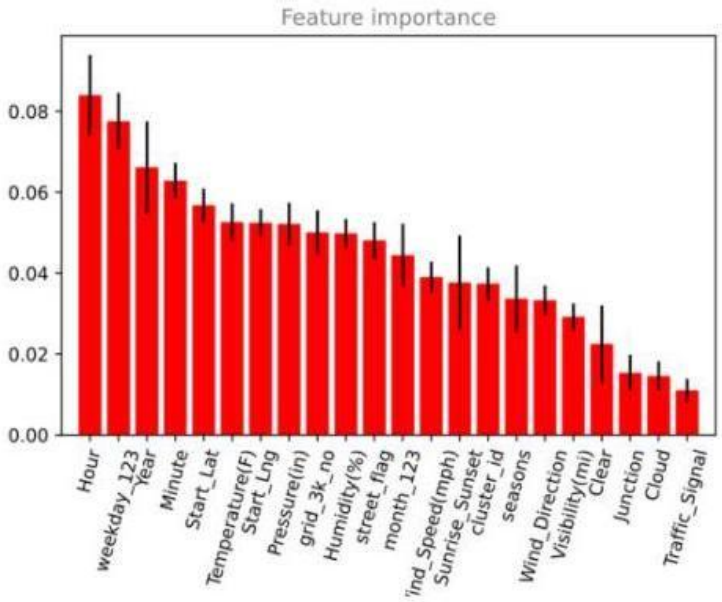


Figure 3: Calculation of Feature Importance Based on ERT

2.3.2 Pearson Correlation Analysis

The strength of the correlation between feature variables and the target variable was analyzed. The Top-23 feature variable set was selected from the feature importance list. The Pearson product-moment correlation coefficient (Pearson's r) with the target variable was calculated, which is the ratio of the covariance to the product of the standard deviations, with the calculation formula as follows:

$$r = \rho_{X_i,Y} = \frac{E(X_i Y) - E(X_i)E(Y)}{\sqrt{E(X_i^2) - E(X_i)^2}\sqrt{E(Y^2) - E(Y)^2}} \quad (1)$$

Figure 4 shows the Pearson correlation calculation results. Overall, the correlation between the target variable Severity and the given feature variables is not high. Among them, there is a relatively strong negative correlation with temporal feature variables such as Year, Month, and Hour; Pressure and Humidity have a relatively strong positive correlation with the target variable; and other feature variables not shown in the figure have no correlation with the feature variables.

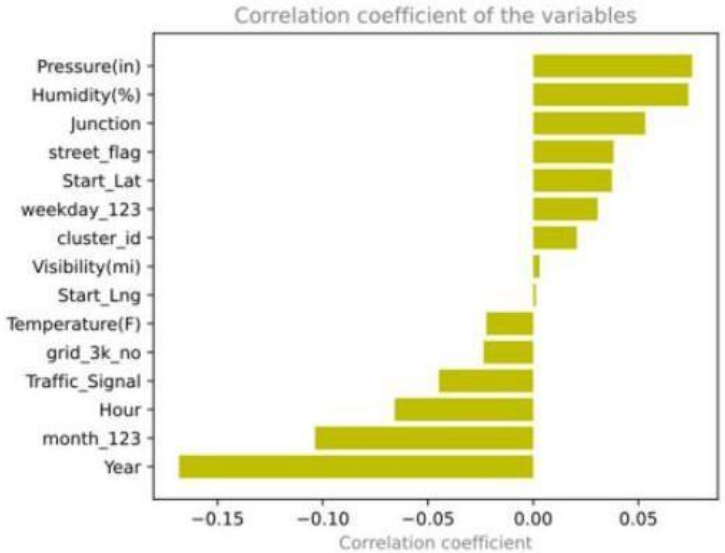


Figure 4: Correlation Between Feature Variables and Target Variable

2.3.3 Statistically-based Association Analysis

(1) Temporal Data Analysis

Statistical analysis of monthly accident volumes from 2017 to 2019 reveals that traffic accidents are most likely to occur in October, followed by December and September, while February and July have the lowest accident volumes each year, as shown in Figure 5.

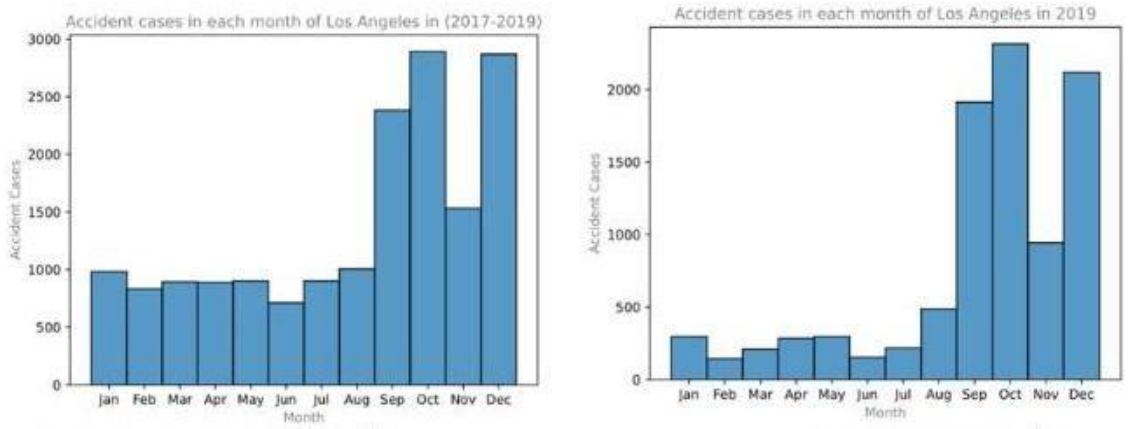


Figure 5: Histogram of Monthly Traffic Accidents

The overall data was aggregated by weekdays and weekends, as shown in Figure 6. The leftmost line chart displays the changes in average accident severity throughout the week, indicating that the average accident severity on weekends is higher than on weekdays; the other two charts respectively show the line and bar statistics of accident volumes for the seven days of the week, with accidents on weekdays being approximately 87% more than on weekends.

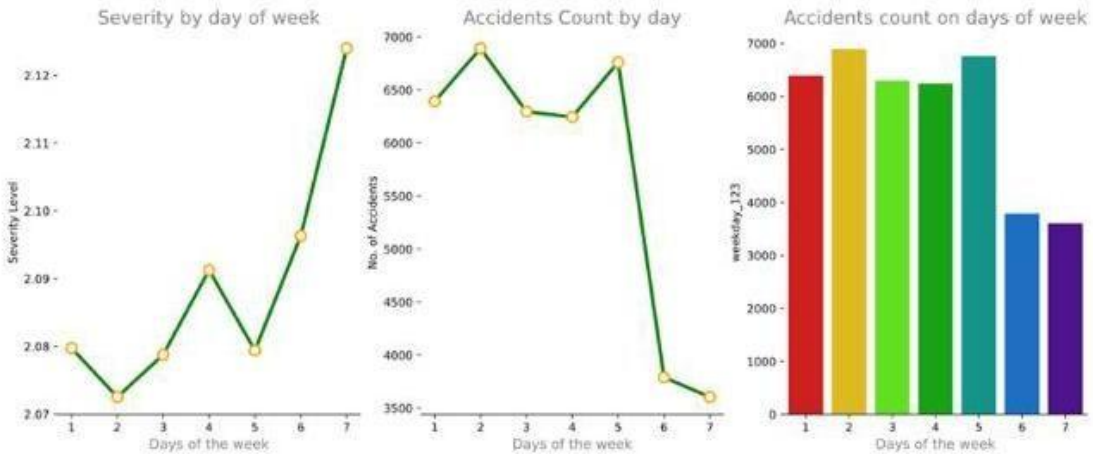


Figure 6: Traffic Accidents on Weekdays and Weekends

Statistics on accident volumes during different time periods on weekdays and weekends are presented in Figure 7. On weekdays, the number of accidents reaches its peak between 14:00 and 20:00. On weekends, the distribution of accidents throughout the day is more scattered, with no obvious peak.

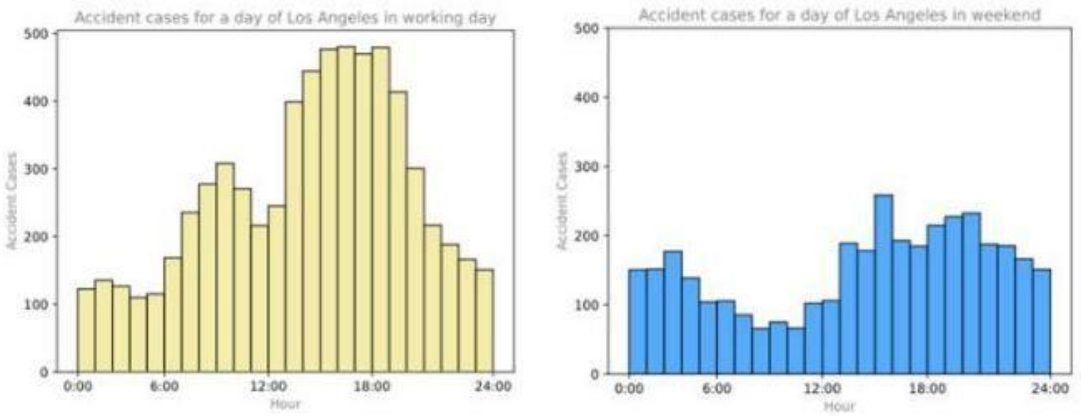


Figure 7: Statistics of Accidents in Different Time Periods

Based on the above temporal correlation analysis results, features such as Month, Week, and Hour were selected as potential temporal prediction features for traffic accidents.

(2) Spatial Data Analysis

Data shows that there are 1095 roads in Los Angeles, with the maximum number of accidents on a single road being 2387, while most roads have fewer than 5 accidents. 34.15% of accidents occur on the Top 10 roads, which account for 1% of the total roads. Figure 8 visualizes the Top 10 roads with the highest traffic accident volumes from 2016 to 2020.

Road hazard grades are classified as shown in Table 4, with the classification standard being the frequency of accidents on the road. Roads with hazard grade 1 have the highest accident frequency and the highest degree of danger. The hazard grade is then constructed as a new attribute "street_flag".

Table 4: Classification of Road Hazard Grades

Danger Level	Level	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10
Number of Accidents	≥2k	≥ 1k	≥ 700	≥ 500	≥ 300	≥ 100	» 39	≥10	≥5	≤ 5

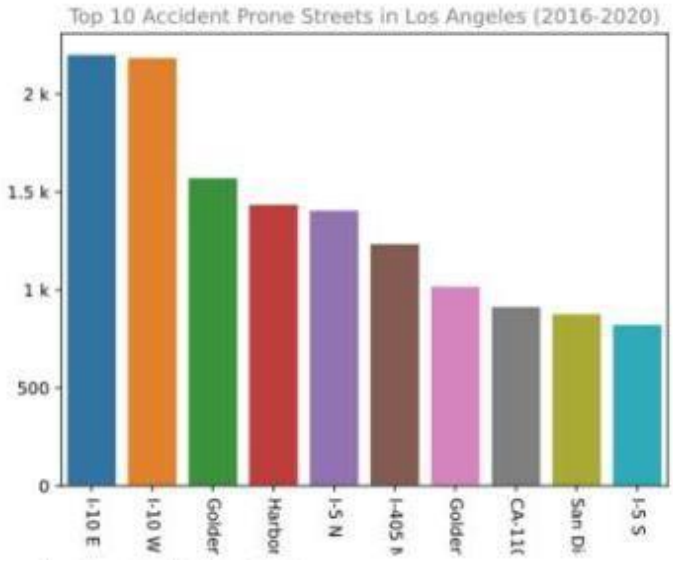


Figure: 8 TOP10 roads by number of accidents

I-10 E: 5.50%
I-10 W: 5.46%
Golden State Fwy S: 3.93%
Harbor Fwy N: 3.59%

I-5 N:3.51%
I-405 N: 3.09%
Golden State Fwy N: 2.54%
CA-110 N: 2.28%
San Diego Fwy S: 2.19%
I-5 S:2.06%
total: 34.15%

Figure 9 shows the situation of missing POI labels at traffic accident locations. It can be found that the vast majority of accident locations lack POI labels; for example, the lack of Bump POI is as high as 99.99%.

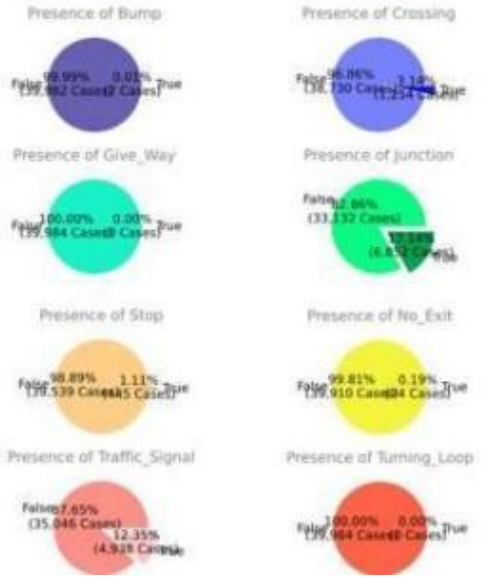
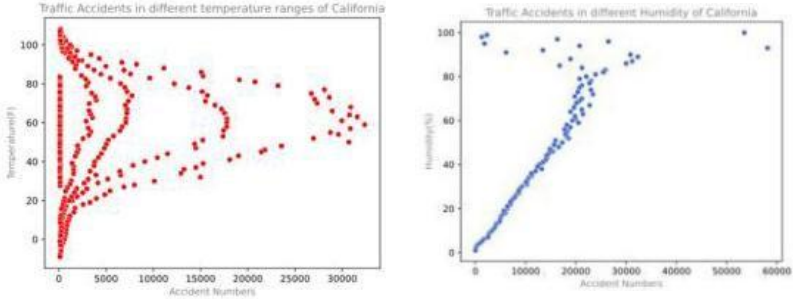


Figure 9: POI missing situation

Based on the above spatial correlation analysis results, POI annotations such as road hazard level (street_flag), presence of traffic signals (Traffic_Signal), presence of junctions (Junction), and the cluster ID (cluster_id) constructed in 3.2.5 are selected as feature inputs, serving as potential spatial prediction features for traffic accidents.

(3) Environmental Data Analysis

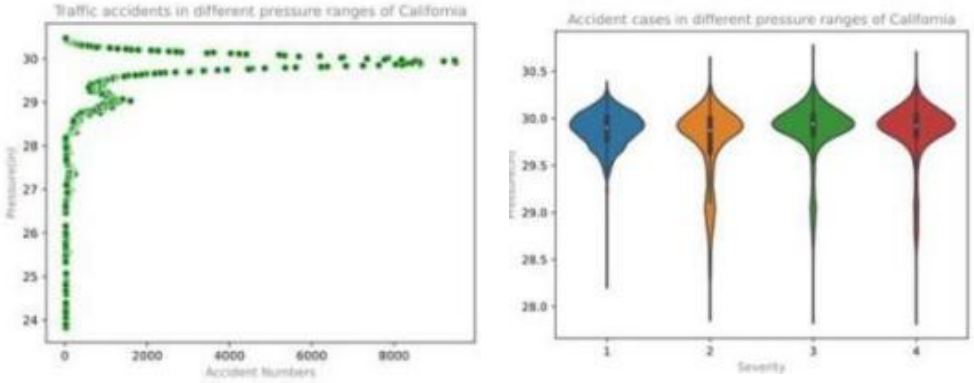
To better study the correlation between temperature, humidity, and accident volume, the data for the entire state of California is represented as scatter plots, as shown in Figure 10. From (a), it can be seen that accidents occur most frequently within the temperature range of 50°F to 70°F, and rarely occur below 20°F or above 90°F. From (b), it can be observed that there is a certain positive relationship between humidity and the number of accidents; as humidity increases, the number of traffic accidents also increases. However, when humidity exceeds 80%, this correlation disappears.



(a) Temperature (b) Humidity

Figure 10: Interrelationship between Temperature, Humidity, and Accident Volume

The correlation between air pressure and traffic accidents was examined using a scatter plot matrix, as shown in Figure 11. From Figure (a), it can be seen that traffic accidents are most likely to occur when air pressure is around 30 inches. Figure (b) displays the overall distribution of traffic accidents of different severity levels under varying air pressures in the form of a violin plot. It can be observed that all traffic accidents of different severity levels are concentrated in the range of 29.5 inches to 30.3 inches. Severity-2 shows a relatively obvious peak at 29 inches, with the lowest median and the largest interquartile range.



(a) Air Pressure Scatter Diagram (b) Air pressure distribution

Figure 11: Interrelationship between Air Pressure and Accident Volume

Based on the above environmental correlation analysis results, environmental factors such as temperature (Temperature), humidity (Humidity), and air pressure (Pressure) are selected as model feature inputs, serving as potential environmental predictors for traffic accidents.

2.3.4 Feature Selection

Based on the above data analysis results, 26 features were selected from the modified dataset for traffic risk prediction from four dimensions: space, time, POI, and environment, as shown in Table 5.

Table 5: Feature Selection

Type	**Feature**			
Space	Start_Lat	Start_Lng	cluster_id	street_flag
Time	Year	month_123	weekday_123	Hour
	Minute	Sunrise_Sunset		
POI	Crossing	Junction	Traffic_Signal	Station
	Stop			
Environment	Temperature(F)	Humidity(%)	Pressure(in)	Visibility(mi)
	Wind_Direction	Wind_Speed(mph)	Cloud	Clear
	Rain			

3. TRAFFIC RISK PREDICTION MODEL

3.1 Overall Algorithm Architecture

The algorithm architecture of the traffic risk prediction model is shown in Figure 12. The model consists of three parts: a "traffic accident binary classifier", an "accident severity multi-classifier", and a "traffic risk assessment function".

Among them, the traffic accident binary classifier is used to predict whether a traffic accident will occur on a certain road section under a certain scenario. During the training phase, the input is the feature variable y extracted from the positive sample set and the Unlabeled sample set, and the output is the probability that the traffic accident is TRUE.

The accident severity multi-classifier is used to determine the severity of a predicted traffic accident in a certain location. The input is the feature variable x obtained from the positive sample set, and the output is the numbers 2, 3, and 4 representing the accident severity.

The traffic risk assessment function evaluates the risk level of a certain block based on the accident occurrence probability and accident severity, and outputs a risk value.

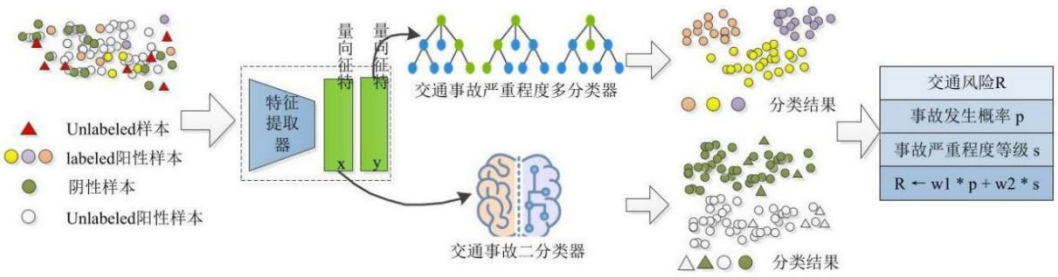


Figure 12: Algorithm Architecture Diagram

3.2 Traffic Accident Binary Classifier

The design process of the traffic accident binary classifier is shown in Figure 13. First, data analysis is performed and feature variables are constructed; then, an Unlabeled sample set is constructed by means of feature recombination, and through PU learning, a credible division of the negative sample set is achieved, on this basis, the training set and test set are further constructed; finally, a binary classification model is built, and the selection of optimal model parameters is realized through k-fold cross-validation and grid search.

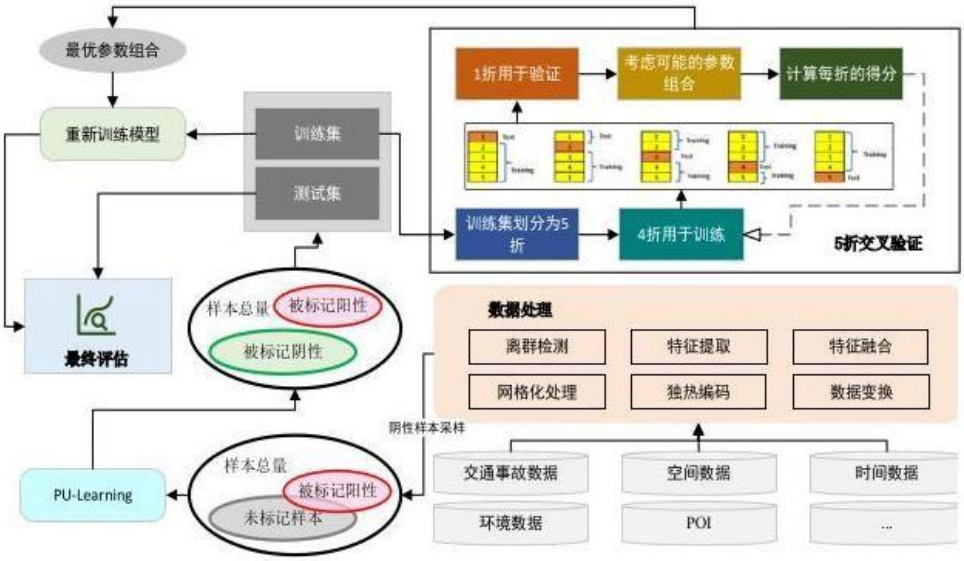


Figure 13: Design Diagram of Traffic Accident Binary Classifier

Construction of Unlabeled Sample Set: All samples in the current dataset are positive data. To judge the probability of a traffic accident occurring in a certain location, the population must include negative samples. Therefore, to train a binary classification model, negative samples first need to be sampled. This paper adopts a negative sample sampling method based on feature recombination, that is, randomly selecting values of different fields in the dataset to generate a record. As long as the record does not appear in the dataset or the negative list, it is added to the negative list. The sample construction process is as follows:

(1) Vertical Data Segmentation

First, according to spatiotemporal sensitivity, the dataset is vertically divided into a Time Sensitive Data-frame (TSD) and a Spatial Sensitive Data-frame (SSD). TSD includes time and corresponding environmental variables

such as temperature and humidity under that time condition, and the other parts are divided into SSD. On the one hand, it can split heterogeneous strongly correlated data; on the other hand, it can well retain the spatiotemporal characteristics of the samples.

(2) Random Sampling

Then, shuffle these two data frames, randomly extract data subsets from the two data frames respectively through random sampling, and combine the data to obtain a new sample.

(3) Reliability Detection

Finally, check if the sample is included in the positive samples. If not, add it to the Unlabeled sample set; otherwise, discard it. The Venn diagram of the relationships between the sample sets is shown in Figure 14.

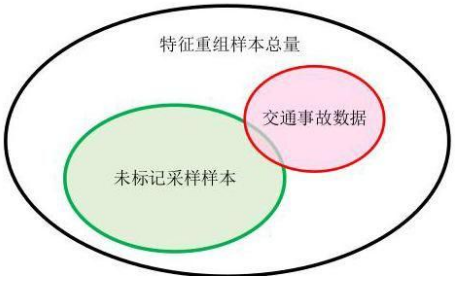


Figure 14: Venn Diagram of Sample Set Relationships

3.3 Multi-classifier for Accident Severity

First, perform oversampling and undersampling on the dataset to address the inherent imbalance of the dataset. Then, combined with cross-validation and grid search methods, the prediction performance of machine learning algorithms such as logistic regression, KNN, decision tree, and random forest was compared, and the random forest model was chosen for implementation.

Figure 15 shows the proportion distribution of traffic accident severity in the sample set. It can be seen that the proportion of Severity-2 samples is as high as 93.2%, far higher than the sample quantities of the other two categories, indicating a significant class imbalance in the data. Creating a traffic severity classification model under such data distribution will result in falsely high classification performance.

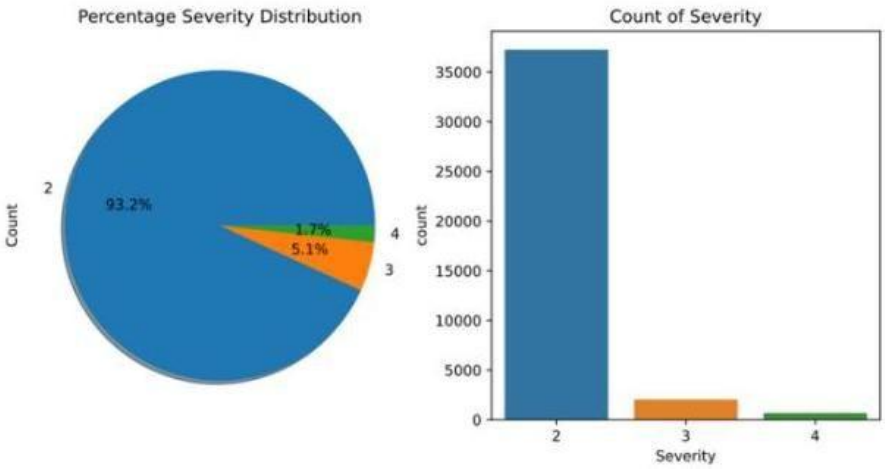


Figure 15: Distribution of the Original Dataset

Randomly undersample the number of Severity-2 samples to 20,000, i.e., randomly select 20,000 samples with Severity equal to 2 from the dataset; oversample the numbers of Severity-3 and Severity-4 to 20,000, i.e.,

randomly select samples with Severity equal to 3 and 4 from the dataset respectively, and replicate them to expand the sample size until the quantity reaches 20,000. The class proportion after sampling is shown in Figure 16.

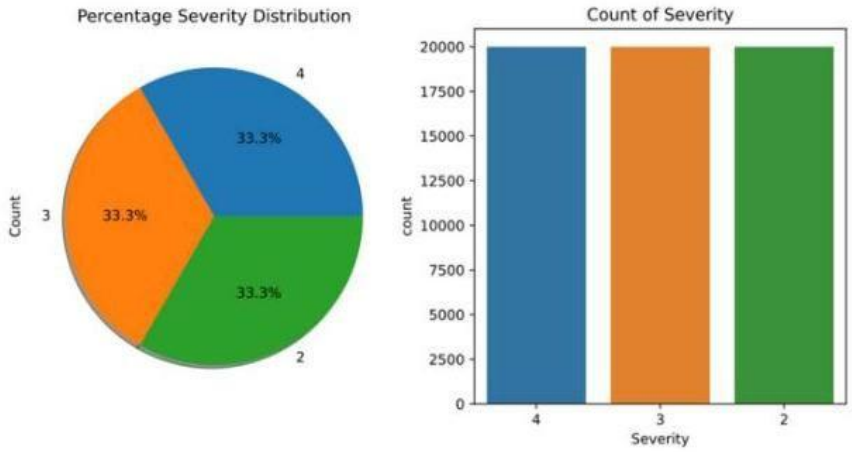


Figure 16: Distribution of the Dataset after Sampling

3.4 Traffic Risk Assessment Function

Combining the probability of accident occurrence and accident severity, the following risk assessment function is constructed, and the meanings of each parameter are shown in Table 6:

$$R(t) = w1 * p(f(\cdot), t) + w2 * s(g(\cdot), t) \quad (2)$$

Table 6: Parameter meaning

Parameter	Meaning
	The magnitude of risk occurring over time
	Feature parameter used to predict the probability of accident occurrence
	Probability of traffic accident occurrence
	Feature parameter used to predict accident severity
	Accident severity level
	Weight parameter, adjusting the influence of probability on risk value
	Weight parameter, adjusting the influence of severity on risk value

3.5 Road Discretization

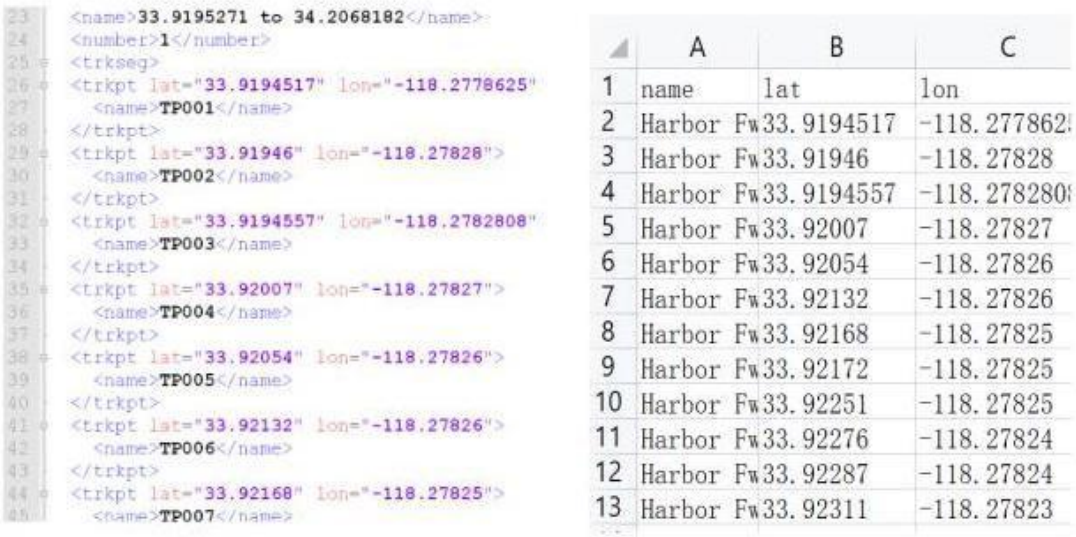
Using the above traffic risk prediction model, real-time risk prediction of urban road traffic accidents can be performed. Each road in the area can be discretized into small road segments of fixed length. At the center of each segment, the model can predict whether a traffic accident will occur by combining given spatiotemporal, environmental and other information, and output the risk value.

(1) Road Information Capture

First, road information for 1095 roads involved in the dataset is captured via the Google Roads API.

(2) Coordinate Point Parsing

The captured Google road information is converted into a GPX file and further parsed into a csv file, as shown in Figure 17 (b), where the "name" column is the road name, and "lat" and "lon" are the latitude and longitude of the path points, respectively. To reduce the model's computational load while ensuring maximum road coverage as much as possible, 1/3 of the parsed data is extracted using uniform sampling.



```
<name>33.9195271 to 34.2068182</name>
<number>1</number>
<trkseg>
<trkpt lat="33.9194517" lon="-118.2778625">
  <name>TP001</name>
</trkpt>
<trkpt lat="33.91946" lon="-118.27828">
  <name>TP002</name>
</trkpt>
<trkpt lat="33.9194557" lon="-118.2782808">
  <name>TP003</name>
</trkpt>
<trkpt lat="33.92007" lon="-118.27827">
  <name>TP004</name>
</trkpt>
<trkpt lat="33.92054" lon="-118.27826">
  <name>TP005</name>
</trkpt>
<trkpt lat="33.92132" lon="-118.27826">
  <name>TP006</name>
</trkpt>
<trkpt lat="33.92168" lon="-118.27825">
  <name>TP007</name>
```

	A	B	C
1	name	lat	lon
2	Harbor Fw	33.9194517	-118.277862
3	Harbor Fw	33.91946	-118.27828
4	Harbor Fw	33.9194557	-118.278280
5	Harbor Fw	33.92007	-118.27827
6	Harbor Fw	33.92054	-118.27826
7	Harbor Fw	33.92132	-118.27826
8	Harbor Fw	33.92168	-118.27825
9	Harbor Fw	33.92172	-118.27825
10	Harbor Fw	33.92251	-118.27825
11	Harbor Fw	33.92276	-118.27824
12	Harbor Fw	33.92287	-118.27824
13	Harbor Fw	33.92311	-118.27823

Figure 17: Coordinate Point Parsing

4. EXPERIMENTAL ANALYSIS

4.1 Performance Analysis of Multi-Classifiers for Accident Severity

After data balancing, machine learning models such as logistic regression, KNN, decision trees, and random forests are used to predict traffic accident severity. 75% of the data is randomly selected as the training set, and the remaining 25% is used for testing. The performance evaluation metrics include Accuracy, Precision, Recall, and F1-Score. Table 7 lists the performance of each model on these evaluation metrics; compared with the benchmark model, the random forest model has the optimal performance.

Table 7: Comparison of Classifier Performance Indicators

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.520	0.520	0.520	0.520
KNN	0.874	0.890	0.870	0.870
Decision Tree	0.915	0.920	0.920	0.910
Random Forest	0.971	0.970	0.970	0.970

Figures 18 and 19 show the ROC curves and PR curves of each model, respectively. It can be seen that the AUC value of the random forest is 0.998, which is higher than that of logistic regression (0.706), KNN (0.968), and decision tree (0.977). This indicates that the random forest model has the best classification effect in terms of the AUC evaluation metric.

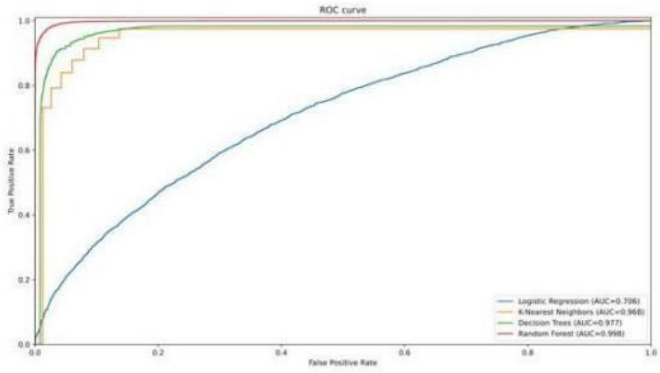


Figure 18: Accident severity classifier ROC curve

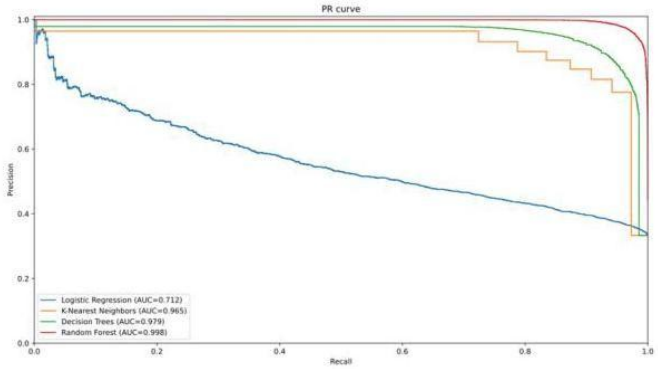


Figure 19: The PR curve of the accident severity classifier

Compared with existing prediction models, Castro et al. [47] used a Bayesian network for accident severity prediction, with a model accuracy of 0.8159, precision of 0.7239, recall of 0.7239, and F1 score of 0.723. Literature [48] divided accident severity into with injured persons and without injured persons, and established a binary classification model for accident severity based on the LightGBM model, with model accuracy, AUC, and F1 score of 0.673, 0.698, and 0.790, respectively; Literature [49] established a three-class RF model for road accident severity, with accuracy, F1 score, and AUC value of 0.853, 0.691, and 0.80, respectively; Literature [50] established an artificial neural network classification model, with accuracy and AUC of 0.746 and 0.752, respectively; Literature [51], which uses the same dataset as this paper, proposed an integrated model RFCNN by combining random forest and convolutional neural network. When all feature variables are used as input, the model accuracy, precision, recall, and F1 score are 0.812, 0.842, 0.864, and 0.853, respectively. After extracting 20 significant features as input, the model has an accuracy of 0.991, precision of 0.974, recall of 0.986, and F1 score of 0.980, with performance slightly higher than this model. However, this model achieves very high model representation ability with lower computational complexity, and the summary comparison is shown in Table 8.

Table 8: Comparison of classifier performance indicators

Model	Accuracy	Precision	Recall	F1-Score	AUC
Bayes [47]	0.816	0.724	0.724	0.723	-
LightGBM[48]	0.673	-	-	0.790	0.698
RF[49]	0.853	-	-	0.691	0.800
Neural Network [50]	0.746	-	-	-	0.752
RFCNN[51]	0.991	0.974	0.986	0.980	-
This Model	0.971	0.970	0.970	0.970	0.998

Ablation experiments were conducted to verify the effectiveness of four attributes, namely "cluster_id", "weekday_123", "grid_3k_no", and "street_flag", constructed through data processing and data analysis, evaluate their impact on model performance, and construct a feature list; then, different subsets were removed from the feature list to generate new feature lists; new random forest classification models were trained, and their performance reports were reviewed.

Through the above verification process, the model comparison and evaluation results are shown in Table 9, with the classification report results retained to three decimal places.

Table 9: Model comparison and evaluation results

Feature List	Accuracy	Precision	Recall	F1-Score
	0.972	0.973	0.972	0.972
	0.964	0.966	0.964	0.964
	0.955	0.959	0.955	0.955
	0.966	0.968	0.966	0.966
	0.967	0.970	0.967	0.967
	0.965	0.967	0.965	0.9654
	0.966	0.967	0.966	0.966
	0.963	0.966	0.963	0.963
	0.954	0.958	0.954	0.954

	0.963	0.965	0.963	0.963
	0.956	0.959	0.956	0.956
	0.950	0.953	0.950	0.949

From the above results, it can be seen that the four constructed attributes all made positive contributions to improving the classification performance of the model; discarding any one or several of these attributes would lead to a decline in model performance. In the worst case, the model's accuracy decreased from 97.2% to 95%, precision from 97.3% to 95.3%, recall from 97.2% to 95%, and F1 score from 97.2% to 94.9%. This ablation experiment confirms the effectiveness of the constructed attributes.

4.2 Performance Analysis of Traffic Accident Classification Model

The model performance report is shown in Figure 20, where "support" indicates the number of samples of the current class in the test set, i.e., the total number of class 0 in the test set is 10003, and class 1 is 9808, with a quantity ratio of approximately 1:1. The model's accuracy, precision, recall, and F1 score are 0.75, 0.74, 0.77, and 0.73, respectively.

[PU Bagging (Random Forest) algorithm] classification_report:

Table10: Traffic accident binary classifier classification report

	Precision	Recall	F1-score	Support
0	0.72	0.78	0.74	10003
1	0.76	0.76	0.72	9808
accuracy			0.75	19811
macro avg	0.74	0.77	0.73	19811
weighted avg	0.74	0.77	0.73	19811

The ROC curve of the model is shown in Fig. 21. The AUC value of the proposed random forest model based on the PU Bagging algorithm is 0.878, which is the highest compared with 8 benchmark models such as logistic regression and KNN. In handling the data imbalance problem, the selection of reliable negative samples is crucial.

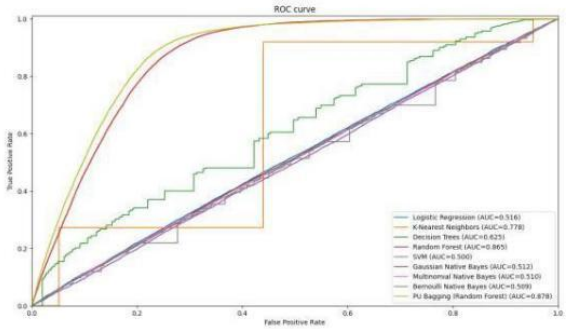


Figure 20: The ROC curve of the traffic accident secondary classifier

Compared with existing models, Reference 52 proposed a binary classification model for real-time traffic accident prediction based on convolutional neural networks, with a prediction accuracy of 0.785; Reference 53 constructed the traffic accident prediction problem as a regression problem and used a deep learning method based on long short-term memory networks to perform real-time risk prediction for specific road sections. This model comprehensively considered influencing factors such as meteorological information, POI, time, and traffic flow, constructed a total of 76 features, and achieved precision, recall, and F1-score of 0.723, 0.773, and 0.736, respectively; Reference 54 built a deep learning model based on recurrent neural networks to predict traffic accident risks, with mean absolute error and root mean square error of 0.014 and 0.034, respectively. However, this model insufficiently analyzes the data affecting accidents, and its prediction granularity is every 3 days, resulting in poor real-time performance.

5. CONCLUSION

Based on historical real traffic accident data, this study deeply explores the intrinsic relationship between road traffic accidents and dynamic influencing factors, comprehensively considers the factors inducing traffic accidents, and improves the predictive ability of the model. Furthermore, a structured real-time traffic risk prediction model is designed, which decomposes traffic risk into two levels: accident occurrence probability and accident severity level. This more meticulously captures the interactions between various factors, enabling these complex relationships to be better modeled and identified, and providing a reliable basis for traffic accident risk prediction. Future research can further optimize the model on this basis and explore more factors affecting traffic safety to continuously improve the scientificity and effectiveness of traffic risk management.

REFERENCES

[1] Tu, Tongwei. "AutoNetTest: A Platform-Aware Framework for Intelligent 5G Network Test Automation and Issue Diagnosis." (2025).

[2] Meng, Q., Wang, J., He, J., & Zhao, S. (2025). Research on Green Warehousing Logistics Site Selection Optimization and Path Planning based on Deep Learning.

[3] Wu, W. (2025). Fault Detection and Prediction in Models: Optimizing Resource Usage in Cloud Infrastructure.

[4] Chen, J. (2025). Efficient and Scalable Data Pipelines: The Core of Data Processing in Gig Economy Platforms.

[5] Yuan, J. (2024, December). Efficient techniques for processing medical texts in legal documents using transformer architecture. In 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC) (pp. 990-993). IEEE.

[6] Li, X., Wang, J., & Zhang, L. (2025). Named entity recognition for smart city data streams: Enhancing visualization and interaction. Authorea Preprints.

[7] Deng, X., & Yang, J. (2025, August). Multi-Layer Defense Strategies and Privacy Preserving Enhancements for Membership Reasoning Attacks in a Federated Learning Framework. In 2025 5th International Conference on Computer Science and Blockchain (CCSB) (pp. 278-282). IEEE.

[8] Lin, Z., Liu, X., Xiang, Y., & Hong, Y. (2025). Modeling multivariate degradation data with dynamic Research on the Design of a Sales Forecasting System Based on Hadoop Big Data Analysis

[9] Yi, X. (2025, October). Real-Time Fair-Exposure Ad Allocation for SMBs and Underserved Creators via Contextual Bandits-with-Knapsacks. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1602-1607).

[10] Tang, Y., Kojima, K., Gotoda, M., Nishikawa, S., Hayashi, S., Koike-Akino, T., ... & Klamkin, J. (2020). Design and Optimization of Shallow-Angle Grating Coupler for Vertical Emission from Indium Phosphide Devices.

[11] Deng, X. (2025). Homomorphic Encryption-Based Data Integrity Verification and Anti-Tampering Mechanism in Cloud Storage Environment.

[12] Mehta, R., Patwar, N., Wei, X., Saunders, E., Zhu, X., & Liu, J. (2026). Towards a National AI Security Framework for Financial Infrastructure Protection. International Journal of Advance in Applied Science Research, 5(2), 39–50. Retrieved from https://h-tsp.com/index.php/ijaasr/article/view/251

[13] Zhou, Z. (2026). Bottleneck Diagnosis in International Automotive Sales Funnels Using Gradient Boosting Trees: Evidence from Cross-Regional Team Efficiency Evaluation. Journal of Computer Technology and Applied Mathematics, 3(1), 11-18.

[14] Wensi, L. (2026). AI-Enabled Data Visualization Marketing for Automated Production Lines: Building Customer Trust and Improving Lead-to-Order Conversion. Academic Journal of Natural Science, 3(1), 8-13.

[15] Li, W. (2026). AI‐Based Prediction and Management of Automation Equipment Lifecycle Costs: A Pathway to Enhancing Customer Lifetime Value (CLV).

[16] Ren, Z. (2024). Adaptive Multi-Scale Fusion for Infrared and Visible Object Detection in YOLOv8. Journal of Theory and Practice of Engineering Science, 4(09), 28–34. https://doi.org/10.53469/jtpes.2024.04(09).04

[17] Ximeng, Y., & Yiming, Z. (2026). Offline Conservative RL for Transaction Authorization: Smartly Balancing Fraud Risk and Customer Friction. Journal of Economic Theory and Business Management, 3(1), 1-9.

[18] Yang, X., & Zhang, Y. (2026). Edge-Enabled Real-Time Fraud Detection for Network Lending Terminals under Low-Latency Constraints. Journal of Computer Technology and Applied Mathematics, 3(1), 55-62.

[19] Wang, Y., Shen, Z., Hu, K., Yang, J., & Li, C. (2025). AI End-to-End Autonomous Driving.

[20] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).