



OPEN ACCESS

EDITED BY

Stefano de Luca,
University of Salerno, Italy

REVIEWED BY

Muhammad Javed,
Shanghai Maritime University, China
Kuo-Chuan Wu,
National Pingtung University, Taiwan

*CORRESPONDENCE

Quan Yu,
✉ yuquan@ncut.edu.cn

RECEIVED 09 July 2025

REVISED 07 January 2026

ACCEPTED 09 January 2026

PUBLISHED 09 February 2026

CITATION

Cheng R, Liu A, Sun X, Liu F, Li N, Wang Y, Yang L and Yu Q (2026) Freeway traffic state classification using vehicle trajectory data. *Front. Future Transp.* 7:1662480. doi: 10.3389/ffutr.2026.1662480

COPYRIGHT

© 2026 Cheng, Liu, Sun, Liu, Li, Wang, Yang and Yu. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Freeway traffic state classification using vehicle trajectory data

Rende Cheng¹, An Liu², Xiaofei Sun³, Fangliang Liu³, Na Li³, Yu Wang³, Lu Yang³ and Quan Yu^{4*}

¹Henan Zhongyuan High-speed Zhengluo Construction Co., Ltd, Zhengzhou, China, ²Jiangxi Communications Investment Group Co., Ltd, Nanchang, China, ³CCCC Highway Consultants Co., Ltd, Beijing, China, ⁴School of Electrical and Control Engineering, North China University of Technology, Beijing, China

This study proposes the FCM-RF-SMOTE framework to resolve the issue of data imbalance in real-time freeway traffic state classification. The framework integrates Fuzzy C-Means (FCM), Random Forest (RF), and the Synthetic Minority Over-sampling Technique (SMOTE). Traffic states are classified into four categories (smooth, stable, congested, and severely congested) based on quantitative thresholds derived from FCM clustering centers. The validation utilizes SUMO simulation with Gaussian noise and a 10 Hz sampling rate to approximate millimeter-wave radar characteristics. Results show that the proposed framework significantly increases the representation of the severe congestion class from 3.67% to 19.83%. Consequently, the overall classification accuracy is enhanced from 77.67% to 97.80%, demonstrating superior performance in handling imbalanced datasets compared to baseline methods. The findings demonstrate the robustness of the algorithm for traffic monitoring systems, particularly in identifying minority traffic states, with future work planned for physical sensor validation.

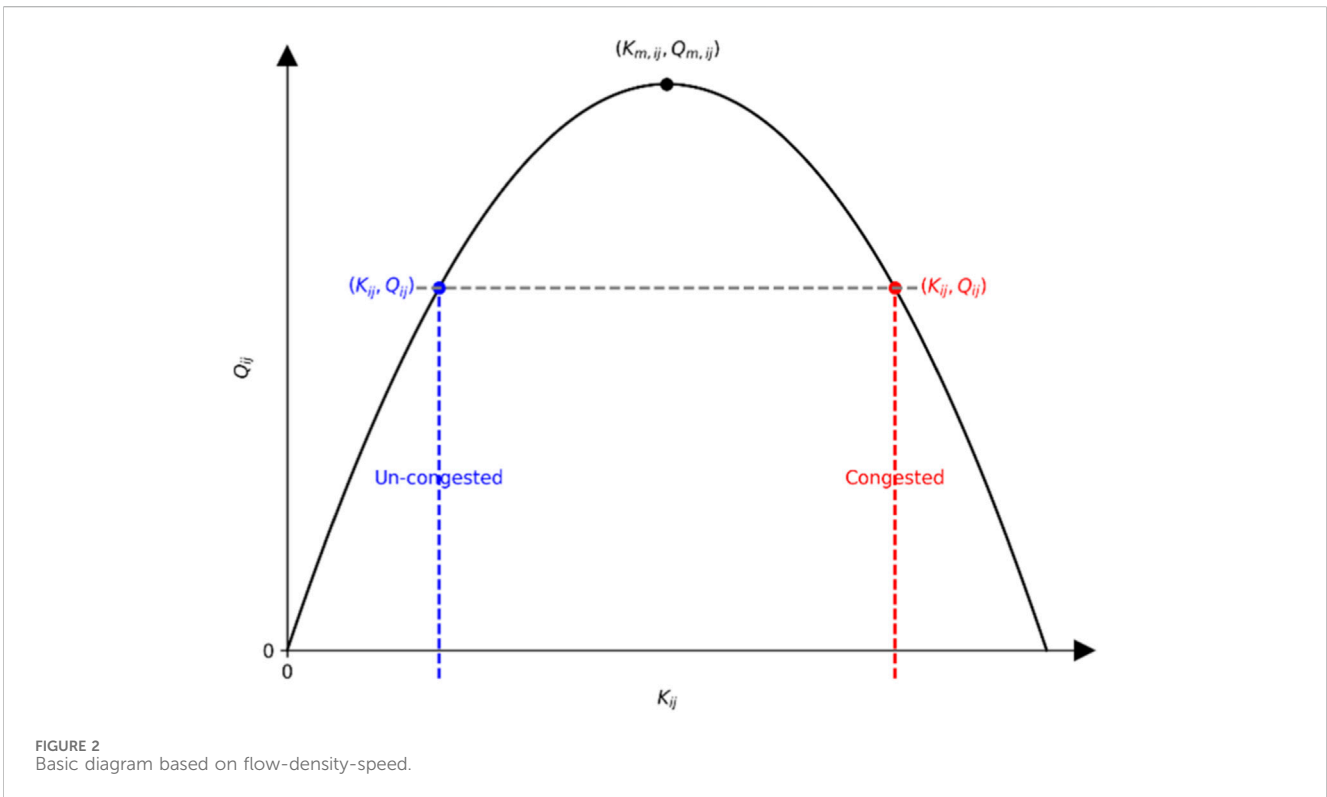
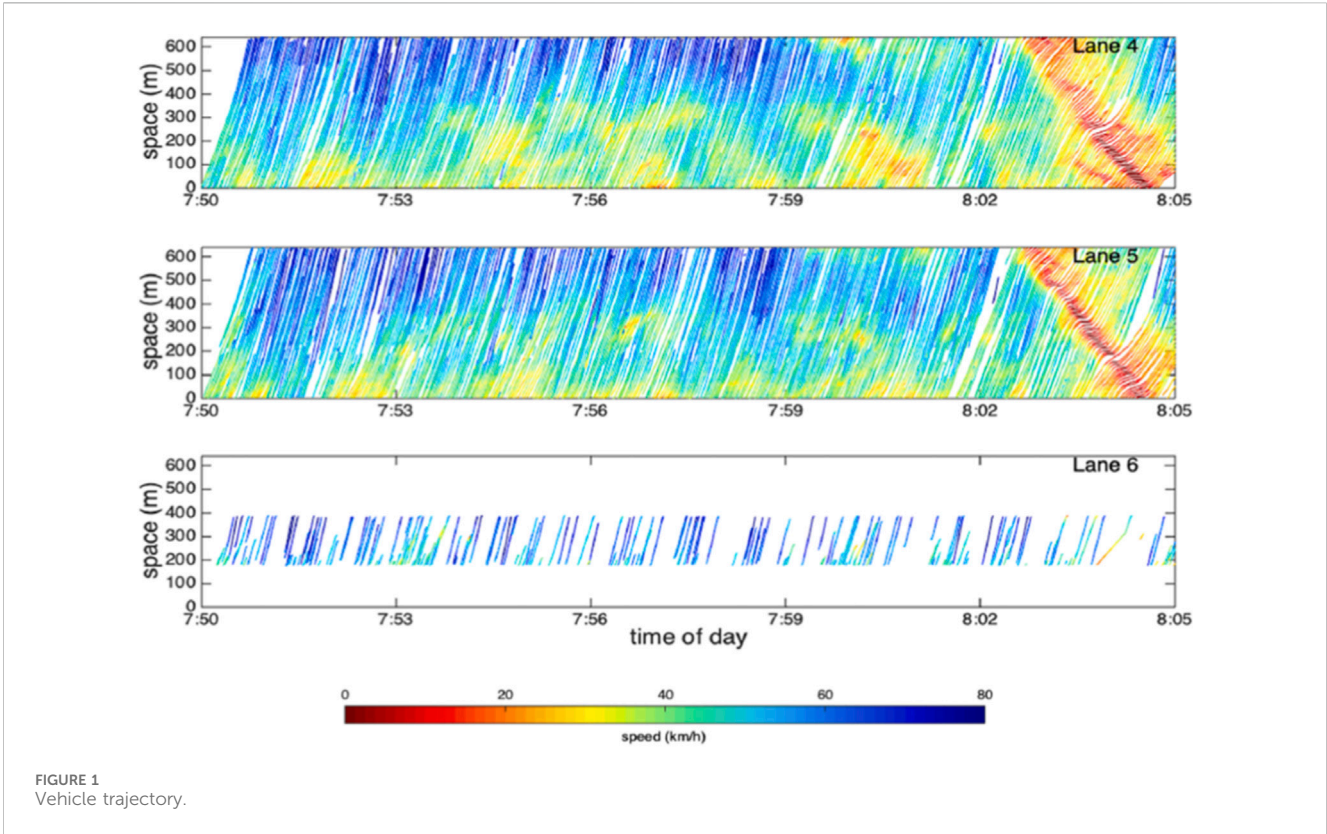
KEYWORDS

freeway, fuzzy c-means, random forest, SMOTE, traffic state classification, vehicle trajectory data

1 Introduction

Traffic condition classification refers to the process of classifying traffic conditions on roads into different categories based on various traffic parameters and data sources. The goal is to provide accurate information for effective traffic management and control. In this study, traffic state classification involves categorizing real-time traffic into four types: smooth, stable, congested, and severely congested. The parameters used for this classification are speed, speed deviation, headway, and density. The thresholds for each category are derived from the FCM clustering center (see Table 4). Specifically, we analyze vehicle trajectory data and extract key parameters, including speed, headway, and density, to accurately identify traffic conditions.

Real-time monitoring of freeway traffic conditions is of utmost importance for traffic managers to effectively manage traffic operations and provide accurate travel guidance, which is an essential component of an active traffic management system. With the development of technology, advanced traffic detectors have emerged. Unlike traditional induction coil detectors, which primarily count traffic volume and occupancy to infer traffic metrics, the latest generation of detectors integrates video and radar technologies, enabling the real-time detection of traffic flow and surrounding environmental conditions, more importantly, these detectors can monitor vehicle trajectories. This allows for the calculation of various detailed traffic parameters, providing a more comprehensive understanding of



traffic states (Barth and Boriboonsomsin, 2008; Yuan, 2020; Zahid et al., 2020; Park et al., 2018). In this study, traffic states are defined based on quantitative thresholds of key parameters, including

average speed, headway, and vehicle density. Specifically, four distinct categories are established: (1) smooth traffic (average speed >110 km/h, density <30 veh/km); (2) stable traffic (average

speed between 80 and 110 km/h, density 30–60 veh/km); (3) congested traffic (average speed between 40 and 80 km/h, density 60–100 veh/km); and (4) severely congested traffic (average speed <40 km/h, density >100 veh/km). This standardized classification ensures consistency and interpretability in subsequent traffic state analysis.

In contrast to traditional induction coil detectors (Nanthawichit et al., 2003; Wang et al., 2018), which primarily count traffic volume and occupancy to infer and compute traffic metrics such as vehicle speed, vehicle length, fleet length, and vehicle type, the new detectors possess the capability to monitor the trajectories of all vehicles on the roadway. This advanced functionality allows for the calculation of headway spacing, headway spacing, speed differences between vehicles, and vehicle density within the detector's coverage area, based on the location and speed of the vehicles. The spatiotemporal velocity distribution of different lanes on a certain road during the morning rush hour is shown in Figure 1, based on the basic graph of traffic density velocity, while the basic graph of traffic density *versus* velocity is presented in Figure 2. These figures provide a comprehensive understanding of vehicle trajectories and their relationship to traffic flow dynamics.

The relationship among flow, speed, and density indicates that the conventional model of traffic states, which relies on the classification of macroscopic fundamental diagrams, fundamentally represents a two-dimensional traffic flow model. In this framework, two parameters are utilized to derive the third parameter. However, advancements in roadside detection technology, which facilitate the acquisition of real-time vehicle trajectory data, allow for an extension of the characterization of traffic operational states from the previous two-dimensional dataset to a higher-dimensional representation.

1.1 Related work

The classification criteria for traffic operation status can be broadly categorized into two types: absolute metrics and relative metrics. Absolute metrics encompass a wide range of fixed values, including traffic volume as outlined in the Freeway Capacity Manual, average travel speed, saturation levels, and the corresponding load factor, which collectively serve as a comprehensive metric. These metrics categorize traffic operating conditions into six classifications, labeled A through F. For instance, China's Interim Technical Requirements for Road Network Operation Monitoring and Services utilize average travel speed as a metric, dividing traffic operating conditions into five categories: smooth, basic smooth, light congestion, moderate congestion, and severe congestion. Such classifications provide quantitative guidelines for assessing traffic states and facilitate the classification of traffic operation conditions through specific metric indicators. However, it is important to note that traffic flow is influenced by a variety of factors, including road characteristics, traffic patterns, weather conditions, and time of day. Additionally, the inherent uncertainty of traffic flow means that a standardized absolute metric may not accurately represent the actual traffic status on basic road segments under varying spatial and temporal conditions (Manual, 2010; Author Anonymous, 2012; Wang, 2019).

These traditional classification criteria have laid a foundation for traffic state analysis. However, with the development of traffic research, more attention has been shifted to traffic state classification algorithms. Early research focused on California and Bayesian algorithms. Notably, in recent years, researchers have made new improvements to these algorithms. Puangnak uses an improved California Algorithm to detect different types of traffic events, effectively improving search capabilities (Puangnak and Chivapreecha, 2019). This improvement broadens the application scope of the California algorithm. However, it may face difficulties in accurately detecting rare or complex traffic events. Shang utilized the Bayesian algorithm to optimize the handling of traffic data imbalance issues (Shang et al., 2021). Jin employed an improved Bayesian algorithm to enhance the classification of relevant traffic flow characteristics (Jin et al., 2023). Zhao applied an improved Bayesian algorithm to the analysis and prediction of road safety conditions (Zhao et al., 2024). Ranpura innovated the traditional Bayesian algorithm by combining real-time traffic data to predict the delay time of traffic vehicles (Ranpura et al., 2024).

This research utilizes various traffic parameters as the foundation for analysis, including traffic flow, speed, and occupancy rate. Traffic state classification was achieved by comparing these parameters against established fixed thresholds (Payne and Tignor, 1978; Cook and Cleveland, 1974; Dudek et al., 1974; Collins et al., 1979; Ahmed and Cook, 1982; Martin et al., 2001; Sheu and Ritchie, 1998). Hsiao et al. (1994) were the first to apply the Fuzzy Logic (FL) algorithm for traffic event classification, employing fuzzy rule formulation and membership functions. Hawas (2007) proposed an urban road traffic event detection algorithm that integrated a fuzzy system to establish the membership function for clustering. Bauza et al. (2010) developed a fuzzy classifier-based method to analyze the traffic state of a road segment by examining vehicle networking, thereby facilitating small-scale traffic state exchanges. Liu et al. (2014) utilized Random Forest (RF) techniques to detect traffic events, effectively addressing noise and overfitting issues. Jiang et al. (2020) employed K-means clustering in combination with a Multi-Layer Perceptron (MLP) to detect urban road traffic status, resulting in a more effective real-time monitoring model.

Recently, Machine Learning methods like RF and Neural Networks have been applied. Sharma employed a neural network based on the Convolutional Neural Network (CNN) model, significantly improving the accuracy of vehicle trajectory prediction on highways (Sharma et al., 2023). Dr. P. Hasitha Reddy utilized a Deep CNN model to analyze traffic monitoring data, which enhanced control capabilities in traffic management (Reddy et al., 2024). Park applied Long Short-Term Memory (LSTM) networks to predict vehicle trajectories and traffic volume on urban roads, offering a potentially more accurate and efficient solution compared to traditional methods (Park and Yoon, 2024). Wan Ming implemented the random forest algorithm to detect traffic violations among taxi drivers and accurately predict the severity of these violations (Ming et al., 2023). This approach demonstrated high efficiency and precision, outperforming conventional detection methods. Shaaban employed the SMOTE to analyze traffic accident data, effectively addressing the issue of data imbalance and enhancing the reliability of subsequent data-based analyses (Shaaban et al., 2024).

TABLE 1 Timetable of traffic flow.

Time(s)	Cars (veh/h)	Trucks (veh/h)
0–3,600	6,300	700
3,600–7,200	1800	200
7,200–10800	4,500	500
10,800–18000	3,600	400
18,000–25200	5,400	600
25,200–32400	4,500	500
32,400–36000	2,700	300

1.2 Problem statement and contributions

Although traffic state classification has evolved from early threshold-based algorithms (e.g., California and Bayesian algorithms) to advanced machine learning methods (e.g., Fuzzy Logic and Random Forest), significant challenges remain. Existing studies predominantly focus on balanced datasets, neglecting the real-world prevalence of data imbalance. For instance, fuzzy logic often relies on subjective membership functions, while standard Random Forest models struggle to accurately classify minority states (e.g., severe congestion) in imbalanced or noisy data environments. Furthermore, while recent deep learning approaches improve prediction, they often overlook sensor noise characteristics.

To address these gaps, this study proposes a novel framework denoted as FCM-RF-SMOTE. This approach utilizes SUMO simulation to generate realistic trajectory data with radar-like noise characteristics. By integrating Fuzzy C-Means (FCM) clustering, Random Forest (RF), and the Synthetic Minority Over-sampling Technique (SMOTE), this study aims to develop a robust model for freeway traffic state classification.

The primary contributions of this study are summarized as follows. A standardized classification system is established. By introducing the SMOTE algorithm, the representation of the minority class (severe congestion) is significantly improved, increasing its proportion from 3.67% to 19.83%. Consequently, the overall classification accuracy enhanced from 77.67% to 97.80%. The fuzzy clustering feature of FCM is utilized to define traffic states objectively, improving upon traditional hard-threshold methods and better reflecting the continuity of traffic flow. The framework is validated using SUMO simulation with Gaussian noise ($\sigma = 0.1$ m) and a 10 Hz sampling rate, demonstrating the algorithm's robustness in scenarios approximating real-world millimeter-wave radar detection.

2 Methodology and experimental design

2.1 Data preparation

In this study, we employ the Simulation of Urban Mobility (SUMO) platform to model vehicle dynamics on a mainline freeway.

The SUMO simulation parameters were calibrated to emulate basic characteristics of millimeter-wave radar. Gaussian positional noise ($\sigma = 0.1$ m) was injected via SUMO 'noise' module, and detectors were configured with a 10 Hz sampling rate to approximate radar measurement intervals. However, this simplified model does not account for multi-target tracking or signal attenuation effects (e.g., rain fade).

The simulation is conducted over a duration of 10 h, utilizing a time step of 1 s and a random seed value of 42. The input traffic flow on the mainline roadway varies between 2000 and 7,000 vehicles per hour. The composition of the vehicle fleet is characterized by 90% small cars and 10% trucks, as detailed in Table 1. Furthermore, to enhance the realism of the simulation, a single accident is introduced at a random point during the simulation period to reflect actual traffic conditions.

The approximate simulation of the characteristics of radar sensors, including noise and multi-target tracking, provides a realistic dataset for model training. The classification performance achieved in this study (overall accuracy >97%) demonstrates the framework's robustness in handling sensor-like noise, which is a common challenge in real-world deployments. This validates the simulation's capability to emulate physical sensor conditions effectively, supporting the feasibility of applying the framework in practical traffic monitoring systems.

It is important to note that while real-world validation is ideal, obtaining high-fidelity trajectory data that includes specific 'severe congestion' events (such as accidents) is extremely difficult and dangerous to instrument in physical environments. Simulation allows us to generate these 'minority class' events safely and consistently, which is essential for validating the effectiveness of the SMOTE algorithm in handling imbalanced data.

The data obtained from the roadside radar video detector encompasses various parameters, including time and coordinate points, vehicle type, instantaneous vehicle speed, and license plate information, among others (see Table 2).

- (1) Speed is calculated using Equation 1

$$V = \frac{\sum_i^i v_i}{i} \quad (1)$$

- (2) The speed deviation for an individual vehicle is defined in Equation 2. The average speed deviation is calculated as shown in Equation 3.

$$vdeviation_i = v_{i+1} - v_i \quad (2)$$

$$V_{deviation} = \frac{\sum_i^i vdeviation_i}{i} \quad (3)$$

- (3) Headway distance is determined by Equation 4. The average headway is calculated using Equation 5.

$$h_i = x_i - x_{i+1} \quad (4)$$

$$H = \frac{\sum_i^i h_i}{i} \quad (5)$$

- (4) Headway time is defined in Equation 6. The average headway time is derived using Equation 7.

TABLE 2 Raw data table.

Time	vehicle_id	vehicle_lane	vehicle_speed	vehicle_type	vehicle_x	vehicle_y
0	E2.1.0	L0_0	20	Car	0	-9.38
0	E2.2.0	L0_1	20	Trucks	0	-5.62
1	E2.1.0	L0_0	22.34	Car	22.34	-9.38
1	E2.1.1	L0_2	20	Car	0	-1.88
1	E2.2.0	L0_1	21.01	Trucks	21.01	-5.62
2	E2.1.0	L0_0	23.78	Car	46.12	-9.38
2	E2.1.1	L0_2	21.43	Car	21.43	-1.88
2	E2.1.2	L0_0	20	Car	0	-9.38
2	E2.1.3	L0_2	20	Car	91.2	-1.88
2	E2.2.0	L0_1	21.99	Trucks	43.01	-5.62
3	E2.1.0	L0_0	26.23	Car	72.35	-9.38
3	E2.1.1	L0_2	23.61	Car	45.04	-1.88
3	E2.1.2	L0_0	21.86	Car	21.86	-9.38
3	E2.1.3	L0_2	22.6	Car	113.8	-1.88
3	E2.1.4	L0_1	20	Car	0	-5.62
3	E2.1.5	L0_1	20	Car	153.51	-5.62
3	E2.2.0	L0_1	23.27	Trucks	66.28	-5.62
4	E2.1.0	L0_0	28.54	Car	100.89	-9.38
4	E2.1.1	L0_2	25.64	Car	70.68	-1.88

TABLE 3 1 min set meter traffic operation parameters.

Time	Speed	Speed_deviation	Headway	Headway_time	Density
660	100.578	4.974465	44.17175	1.583793	68.07377
720	108.2678	3.951056	50.09159	1.677349	57.09016
780	104.741	3.813028	47.10683	1.627219	61.02459
840	107.175	4.493562	50.78096	1.706486	53.27869
900	106.4732	4.300598	46.88246	1.590783	57.7459
960	103.2073	3.826104	44.22546	1.549328	66.96721
1,020	104.3925	5.74591	50.04495	1.726976	56.18852
1,080	111.6654	2.726264	50.88868	1.64871	53.52459
...
35,940	114.7259	5.277375	73.25056	2.265644	27.37705
36,000	115.768	5.394076	71.08095	2.189484	27.08333

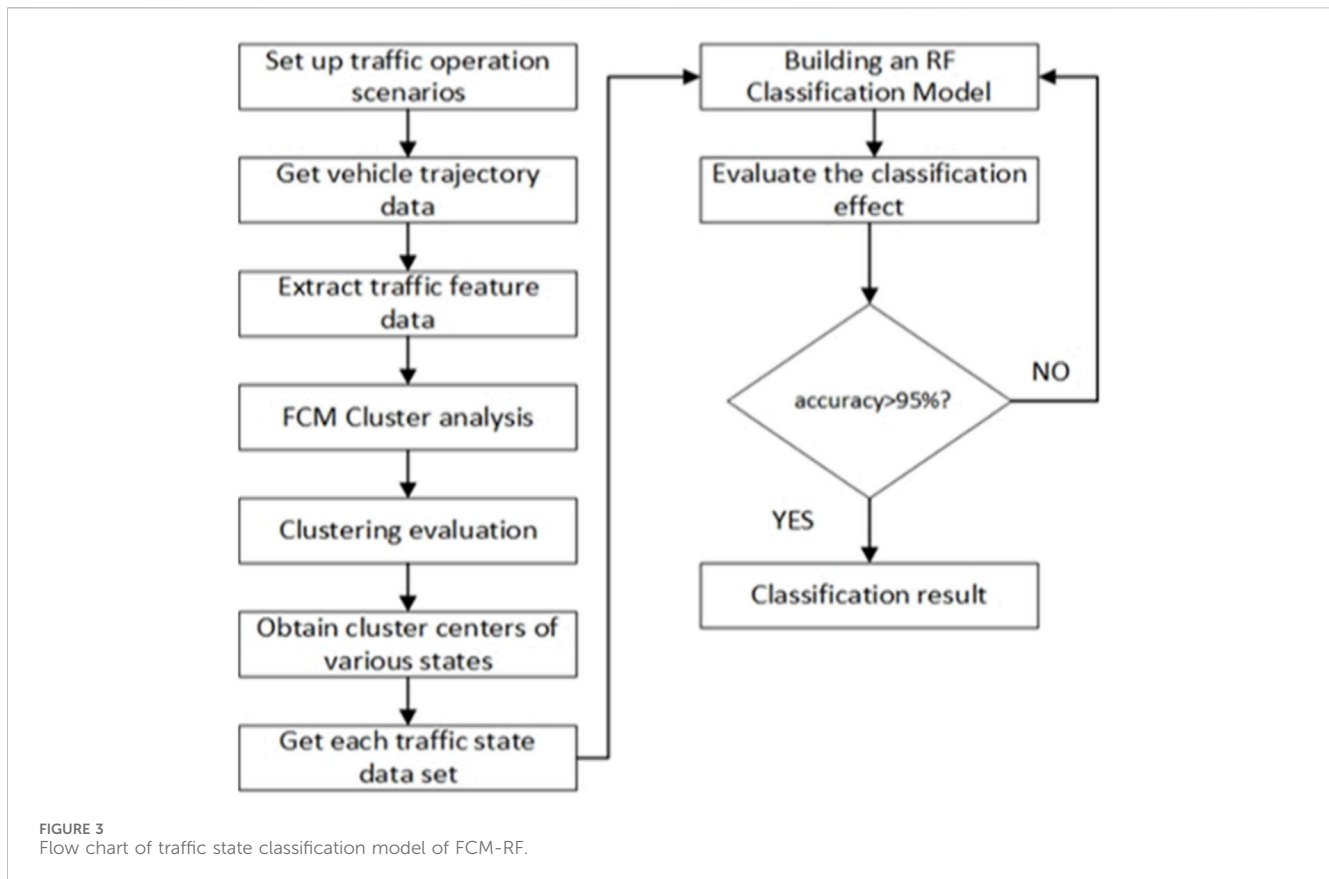
$$ht_i = \frac{x_i - x_{i+1}}{v_{i+1}} \tag{6}$$

$$HT = \sum_i^i ht_i \tag{7}$$

(5) Vehicle density is calculated according to Equation 8.

$$D = \frac{i_{car} + i_{truck} * 1.5}{200} \times 5 \tag{8}$$

Within each 1-s interval, vehicles are organized by lane and direction. Due to high-frequency fluctuations, traffic parameters are aggregated into 1-min intervals to enhance stability. Table 3 presents



the processed dataset parameters: front and rear vehicle positions (x_i, x_{i+1}) and speeds (v_i, v_{i+1}); average interval speed (v); average speed deviation ($V_{deviation}$); average headway distance (H); average headway time (HT); and vehicle density (D , calculated as equivalent passenger cars per kilometer). These aggregated metrics serve as the foundational feature vectors for the subsequent classification model.

2.2 Proposed framework

In contemporary traffic state analysis, addressing data complexities is of paramount importance. The Synthetic Minority Over-Sampling Technique (SMOTE) plays a pivotal role in the preprocessing stage. In traffic datasets, class imbalance is a prevalent issue, characterized by the underrepresentation of certain traffic states. SMOTE addresses this concern by generating synthetic samples for minority classes. By incorporating these synthetic samples into the dataset, SMOTE effectively rectifies the class distribution, ensuring that subsequent classification algorithms are not biased towards majority classes.

Following this, the Fuzzy C-means algorithm is applied to categorize traffic states. In the context of traffic analysis, it employs a set of traffic-status metrics to partition data into distinct traffic-state classes. Metrics such as vehicle speed, traffic flow rate, and occupancy are typically considered. By iteratively updating the membership values of each data point across different clusters and adjusting the cluster centers, the algorithm generates a traffic-status dataset for each class.

Finally, the Random Forest (RF) algorithm is implemented within the traffic-state classification decision module. RF, a combinatorial classifier, operates as a non-parametric classification algorithm. By analyzing the four classes of traffic-state data generated by the Fuzzy C-means algorithm, a traffic-state classifier is developed. This classifier can accurately assess the real-time operational status of road traffic, even in the presence of noisy data and missing values. In order to assess the temporal variability of traffic state changes, detector deployment intervals of 100 m, 200 m, 500 m, and 1,000 m were utilized. In summary, this paper constructs a traffic state discriminative model based on FCM-RF. The process of the FCM-RF traffic state classification model is shown in Figure 3.

2.3 Algorithm principles

2.3.1 SMOTE

The synthetic few oversampling technique (SMOTE) proposed by Chawla et al. generates synthetic samples for minority classes to solve the problem of data imbalance (Chawla et al., 2002). The generation mechanism of this algorithm is simple and mainly consists of two parts: selecting k nearest neighbors of minority class samples based on the measurement method and generating new samples by interpolating between these minority class samples and neighboring samples using a linear interpolation strategy. The specific calculation process is as follows. For each minority class sample x_i , SMOTE:

Finds its k -nearest neighbors;
 Randomly selects one neighbor x_{nm} ;
 Generates a synthetic sample x_{new} along the line segment between x_i and x_{nm} .

A synthetic sample x^{new} is generated along the line segment as shown in Equation 9.

$$x_{new} = x_i + \lambda (x_{nm} - x_i) \quad (9)$$

where λ is a random number between 0 and 1.

2.3.2 Fuzzy C-means

Fuzzy clustering is a type of soft clustering that differs from hard clustering in that the affiliation function value of a sample to each category can vary between 0 and 1. This feature not only highlights the interrelationships among data points but also reflects the potential transitional states between categories. The objective function to minimize is given by Equation 10:

$$J = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \|x_i - v_j\|^2 \quad (10)$$

where:

n is the number of data points; c is the number of clusters; x_{ij} is the membership value of data point x_i to cluster j ; m is the fuzziness parameter (controls the degree of overlap between clusters); v_j is the centroid of cluster j .

The membership values u_{ij} are updated iteratively using Equation 11:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}} \quad (11)$$

Here, $\|x_i - v_j\|$ represents the distance between the i sample and the j cluster center.

The cluster centroids v_j are updated as shown in Equation 12:

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad (12)$$

Fuzzy clustering provides a more objective and realistic representation of data categorization. Clustering analysis indicates that traffic states exhibit an inherent fuzzy nature, characterized by the lack of distinct boundaries between different traffic states. Therefore, this paper proposes the utilization of fuzzy clustering analysis as a methodological approach for evaluating traffic state metrics in basic freeway sections.

2.3.3 Random forest

The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to improve the accuracy and stability of the model. Here is a detailed description of its principle:

The algorithm starts with performing bootstrap sampling on the original training dataset D . For each iteration, a new training subset D_i is randomly sampled with replacement from D . The probability of each sample being selected in each drawing remains the same.

For each bootstrap sample D_i , a decision tree T_i is constructed. During the construction of the decision tree, at

each node, a random subset of features is selected. The best split among these features is chosen based on a certain impurity measure. Commonly used impurity measures include Gini impurity and entropy.

For a node N with K classes and p_k being the proportion of samples in class k in node N , the Gini impurity is defined as $Gini(N) = 1 - \sum_{k=1}^K p_k^2$. The feature that leads to the greatest reduction in Gini impurity is chosen as the splitting feature.

The entropy of a node N is defined as $H(N) = -\sum_{k=1}^K p_k \log p_k$.

Similar to the Gini impurity, the feature that causes the largest decrease in entropy is selected for splitting. The tree is grown until a certain stopping criterion is met, such as a maximum depth d_{max} is reached or the number of samples in a node is less than a certain threshold n_{min} .

Multiple decision trees $\{T_1, T_2, \dots, T_n\}$ are trained in this way, and the Random Forest F is formed by combining these trees. When making predictions, for a classification task, each decision tree in the forest votes for a class. Let $y_i(x)$ be predicted class of the i tree for input x . The final prediction $\hat{y}(x)$ is the class with the most votes, that is, $\hat{y}(x) = \arg \max_k \sum_{i=1}^n I(y_i(x) = k)$. Where $I(\cdot)$ is the indicator function. For a regression task, the average of the predictions of all the trees is usually taken as the final prediction. If $\hat{f}_i(x)$ is the prediction of the i tree for input x , then the final prediction $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \hat{f}_i(x)$.

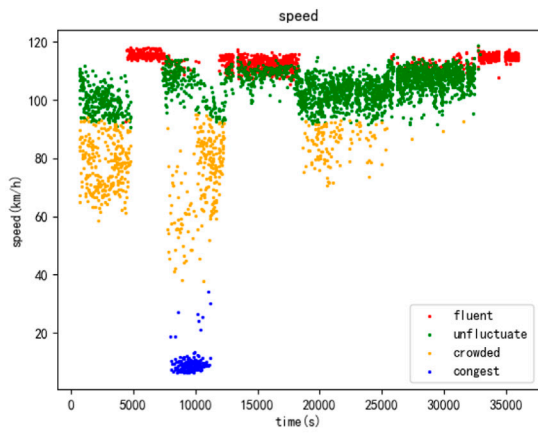
The Random Forest algorithm can also calculate the importance of each feature. One common method is the Mean Decrease in Impurity (MDI). For a feature j , the MDI is calculated as the average decrease in impurity across all trees in the forest when splitting on feature j . Let $I(N)$ be the impurity of node N . The importance of feature j is given by Equation 13:

$$Importance(j) = \frac{1}{n} \sum_{i=1}^n \sum_{N \in T_i} (I(N) - I(N_{left}) - I(N_{right})) \times w(N) \quad (13)$$

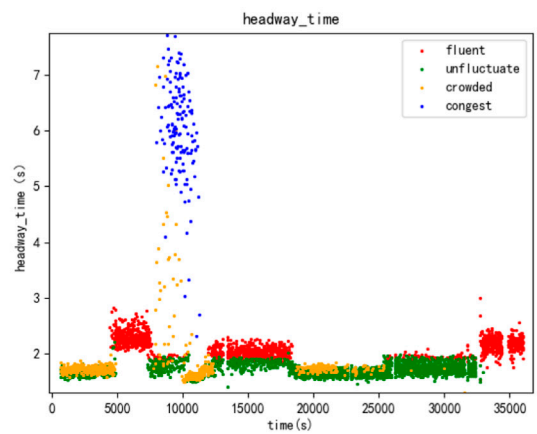
where N_{left} and N_{right} are the left and right child nodes of node N after splitting on feature j , and $w(N)$ is the weight of node N , usually proportional to the number of samples in the node.

Note that all validations in this study are based on simulated data. While the simulation incorporates basic sensor characteristics (noise and sampling rate), it does not account for complex real-world factors like multi-path interference or weather effects, which may impact actual deployment performance.

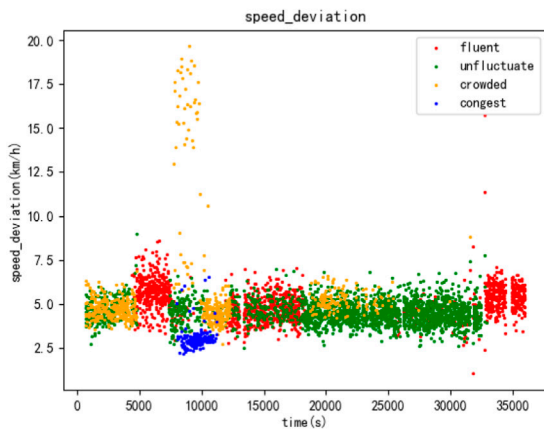
In this framework, the Fuzzy C-Means (FCM) algorithm is employed solely to generate class labels (Smooth, Stable, Congested, Severely Congested) for the training dataset. The input features fed into the Random Forest classifier are the five original traffic parameters extracted from the vehicle trajectory data: interval speed (x_1), speed deviation (x_2), headway (x_3), headway time (x_4), and vehicle density (x_5). The FCM membership values are utilized to determine the hard label for each sample but are not included as feature vectors in the Random Forest training process.



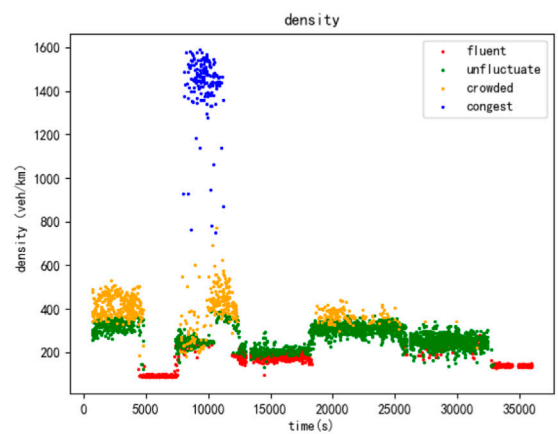
(a) Speed Cluster Picture



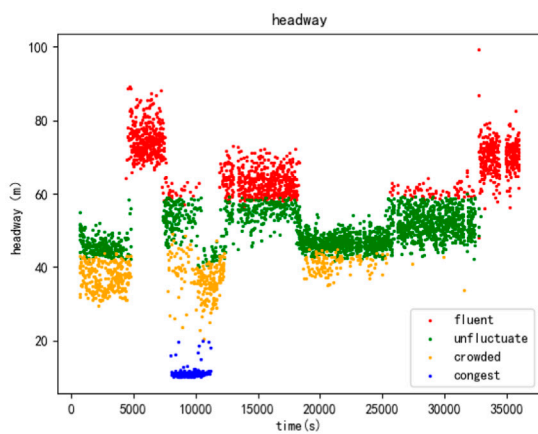
(d) Headway Time Cluster Picture



(b) Speed Deviation Cluster Picture



(e) Density Cluster Picture



(c) Headway Cluster Picture

FIGURE 4 Clustering results for each traffic parameter. (a) Speed Cluster Picture. (b) Speed Deviation Cluster Picture. (c) Headway Cluster Picture. (d) Headway Time Cluster Picture. (e) Density Cluster Picture.

TABLE 4 Clustering center of each traffic state.

State	Speed	Speed deviation	Headway	Headway time	Density	Proportion
Smooth	113.9	5.1	68.0	2.1	29.9	30.86%
Stable	104.7	4.5	49.7	1.7	55.0	51.59%
Congested	77.6	5.5	37.6	2.2	80.9	13.85%
Severely congested	9.7	3.1	11.2	6.3	285.7	3.67%

3 Traffic state characterization based on FCM

Initially, the interval traffic data, organized in 1-min increments, undergoes clustering to classify the various traffic states. Each identified category is assigned a label, thereby generating a training dataset for the state classification algorithm.

The processed feature variables function as training samples, with each sample point represented as $(x_1, x_2, x_3, x_4, x_5)$, where x_1, x_2, x_3, x_4, x_5 correspond to interval speed, speed difference, headway, headway time, and vehicle density, respectively. The Fuzzy C-Means (FCM) algorithm is utilized to cluster the traffic states, identify the category to which each sample belongs, and assign category labels to the samples. Subsequently, the traffic flow parameter variable features of each category are analyzed to characterize specific traffic states.

As illustrated in Figure 4, the traffic flow data have been categorized using the Fuzzy C-Means (FCM) algorithm, resulting in the classification of four distinct clusters that correspond to four categories of traffic states: smooth, stable, congested, and severely congested. Figures 4a,c,e depict the distribution of speed, headway, and vehicle density across these categories, clearly indicating that the smooth state is characterized by the highest speed, significant headway spacing, and low vehicle density. Transitioning from the clear state to the severely congested state is associated with an increase in vehicle density, a decrease in speed, and a further increase in density. Figures 4b,d illustrate the distribution of speed differences and headway spacing across the various categories. In the smooth, stable, and congested states, both speed differences and headway spacing exhibit abrupt changes primarily in response to blockages caused by accidents, while remaining relatively stable in other states. As shown in Table 4, during the detection process, the steady state predominates, accounting for more than 50% of the observations, whereas the severe congestion is the least prevalent, comprising only 3.67%. The variations in each traffic flow parameter align with the flow changes of the simulation input and adhere to established traffic flow theory.

It is worth noting that the state boundaries derived via FCM are data-driven and reflect the intrinsic distribution of the dataset rather than pre-defined engineering standards. As shown in Table 4, the 'Stable' state exhibits higher average speeds (approx. 104.7 km/h) compared to typical urban traffic management thresholds (e.g., 40–60 km/h). This is because the simulation represents a freeway scenario where traffic flow remains fast and stable until it reaches a critical density, after which it rapidly breaks down into congestion. Unlike hard-threshold methods (such as those in the Highway Capacity Manual) which may exhibit subjectivity, the FCM-derived thresholds objectively capture these specific flow-density

transitions inherent to the monitored roadway section. Furthermore, the clustering outcomes for key traffic parameters (speed, headway, and density) align with the findings of Yu et al. (2015), supporting the validity of this approach. Consequently, it is concluded that the traffic states classified by the Fuzzy C-Means algorithm are consistent with the operational characteristics of traffic flow.

4 Classification model results and discussion

4.1 Baseline performance on imbalanced data

The historical traffic data were clustered and analyzed to generate a dataset comprising four distinct types of traffic states. Sixty percent of the data were designated as the training set, while the remaining forty percent were utilized as test samples. The composition of the data samples is presented in Table 5.

The Random Forest (RF) algorithm is executed using the scikit-learn library in Python, and it computes the confusion matrix, which is presented in Table 6.

As shown in Table 7, the accuracy of traffic status classification exceeds 95% when traffic conditions are smooth and stable, indicating the feasibility of the algorithm. However, the limited sample size for congestion and blockage states results in an unbalanced data distribution, which adversely affects the algorithm's discriminatory accuracy. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to ensure a balanced sample size across all categories.

4.2 Data augmentation using SMOTE

The unbalanced data samples from each state adversely affect the performance of the classification algorithm, the distribution of traffic conditions is shown in Figure 5. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic data, which is then incorporated into the original dataset. This approach aims to balance the data distribution and enhance the accuracy of the classification algorithm.

As illustrated in the figure, the distribution of each traffic state is markedly uneven, with the sample size for the stable state significantly exceeding that of the blocked and congested states. This issue of data imbalance leads to erroneous learning by the random forest algorithm, resulting in a low accuracy rate for the blocked and congested states, which subsequently undermines the

TABLE 5 Sample composition of the training and test sets of the algorithm.

Dataset type	Smooth	Stable	Congested	Severely congested	Total sample
Training set	821	1,372	368	98	2,659
Test set	547	915	246	66	1774
Total	1,368	2,287	614	164	4,433

TABLE 6 Classification algorithm confusion matrix.

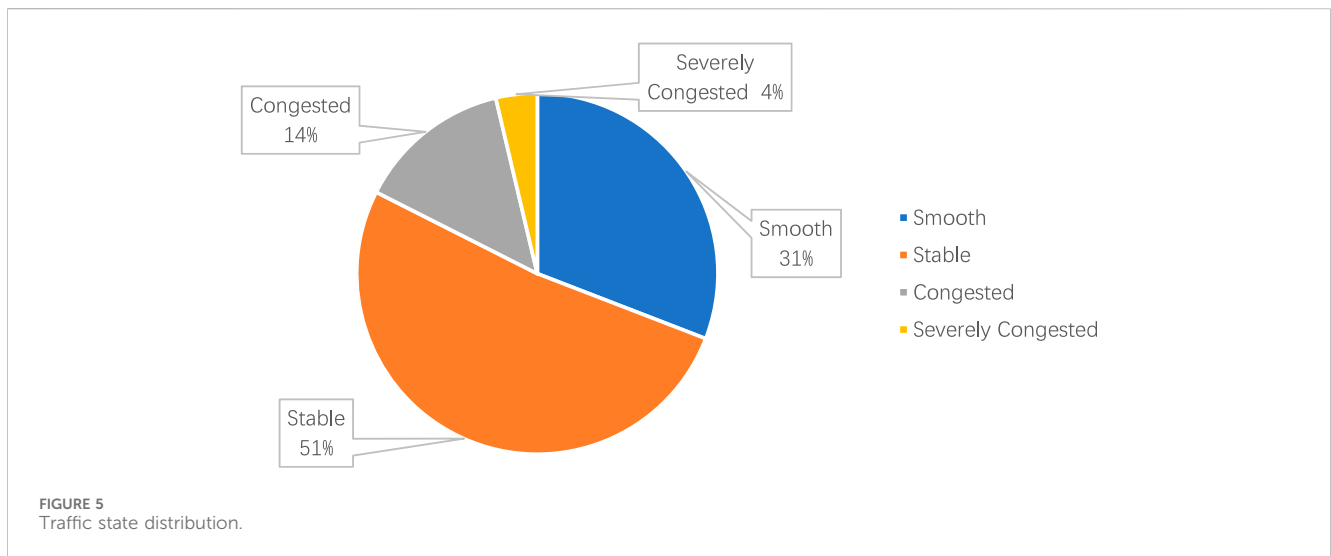
Actual state		Predict			
		Smooth	Stable	Congested	Severely congested
Actual	Smooth	532	0	11	0
	Stable	15	892	0	6
	Congested	0	6	247	65
	Severely congested	0	0	0	0

TABLE 7 Random forest discriminant accuracy rate.

Traffic state	Precision
Smooth	0.9797
Stable	0.9769
Congested	0.7767
Severely congested	0

TABLE 8 Statistical distribution of traffic status data.

Traffic state	Number	Proportion
Smooth	1,368	30.86%
Stable	2,287	51.59%
Congested	614	13.85%
Severely congested	164	3.67%



effectiveness of traffic state classification. Consequently, the application of the SMOTE algorithm is necessary to address the imbalance in the data.

From Table 8, it can be observed that after clustering the traffic flow dataset into 1-min intervals, there are only 164 instances of

severe congestion and 614 instances of congested state. By employing the SMOTE algorithm to synthetically augment the blocked data, the quantity of severe congestion data has been increased sevenfold, while the quantity of congested state data has been doubled. The distribution of data categories for each

TABLE 9 Distribution of traffic states after SMOTE.

Traffic state	Number	Proportion
Smooth	1,368	22.08%
Stable	2,287	36.91%
Congested	1,312	21.18%
Severely congested	1,228	19.83%
Total	6,195	100%

TABLE 11 Random forest discriminant accuracy rate.

Traffic state	Precision
Smooth	0.9797
Stable	0.9856
Congested	0.9823
Severely congested	0.9780

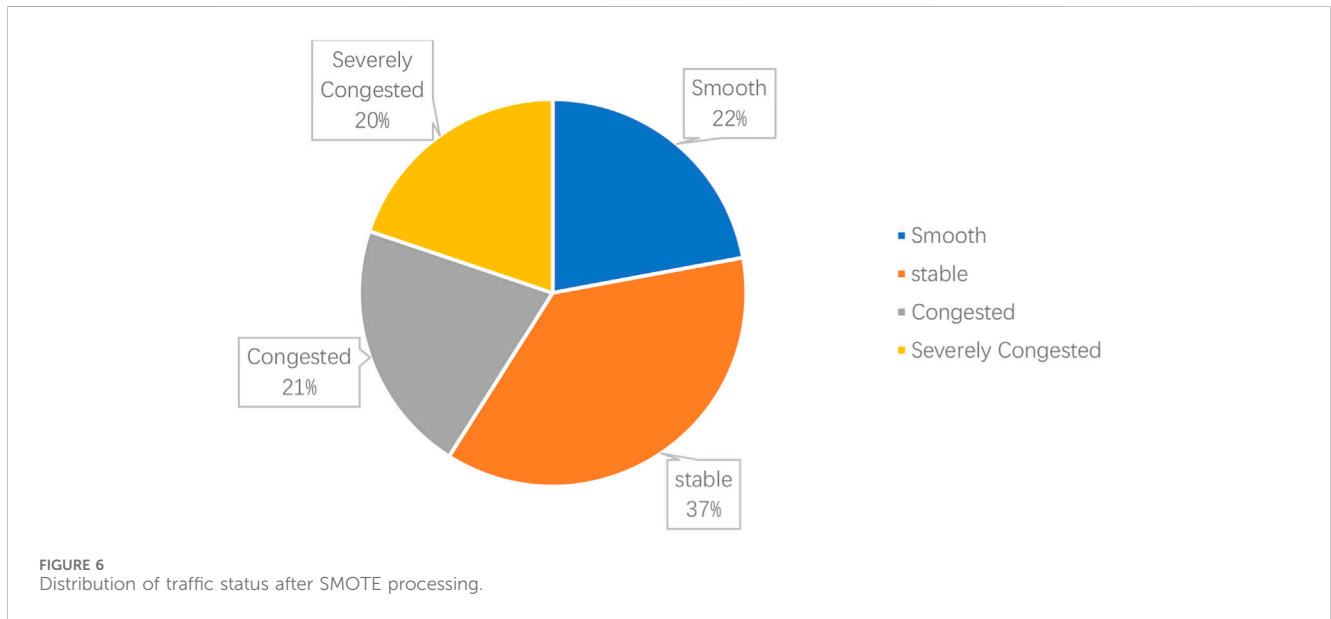


TABLE 10 The confusion matrix of the algorithm.

Actual state		Predict			
		Smooth	Stable	Congested	Severely congested
Actual	Smooth	530	0	17	0
	Stable	8	900	7	0
	Congested	0	0	491	0
	Severely congested	0	5	0	520

state following the application of the SMOTE algorithm is presented in Table 9 and illustrated in Figure 6.

The original dataset has been enhanced with data on congested and severely congested states to ensure that the proportions of the four traffic states are balanced, while also preserving a larger sample size for the stable state. After applying the SMOTE algorithm, the distribution of the four traffic states has become more balanced, with severe congestion now accounting for 19.83% of the total. This adjustment effectively addresses the class imbalance issue present in the original dataset, where severe congestion only made up 3.67%. As a result, the dataset processed using the SMOTE algorithm can now serve as a more representative sample for traffic state classification.

4.3 Performance of FCM-RF-SMOTE framework

Using the balanced dataset generated by SMOTE (as detailed in Table 9), the Random Forest model was retrained using the same 60:40 training-test split ratio. The Random Forest (RF) algorithm is implemented using the scikit-learn library in Python, and it computes the confusion matrix, as illustrated in Table 10.

As illustrated in Table 9, the accuracy of the four categories of traffic state classification exceeds 97%, indicating superior algorithm performance. The application of SMOTE (Synthetic Minority Over-sampling Technique) to address imbalanced data has enhanced the

discriminative capability for congestion and blockage states. As illustrated in Table 11, the classification accuracy exceeds 97% in simulation, indicating potential for real-world applications pending field validation.

4.4 Discussion

The comparative analysis between the baseline and the proposed framework demonstrates the critical role of data balancing. While the baseline RF model struggled with minority classes due to data imbalance (0% precision for severe congestion), the integration of SMOTE significantly enhanced the model's sensitivity, boosting the classification accuracy for severe congestion to 97.80%. These results demonstrate the framework's robustness, which is further validated by the rigorous design of the simulation environment. The emulation methodology is informed by established research in vehicular sensing and communications, notably the contributions of Sommer (Sommer et al., 2011) et al., who elucidated the effects of shadowing and signal degradation on vehicular communications, and Stiller (Stiller et al., 2025), who conducted a review of sensor fusion techniques aimed at enhancing the reliability of traffic monitoring. Furthermore, insights from Li and Yoon (Li and Yoon, 2023) regarding radar-camera fusion have further guided the design of the simulation, effectively replicating sensor imperfections and multi-target scenarios that are typical in real-world applications.

The use of SUMO simulation allowed for the injection of Gaussian noise ($\sigma = 0.1$ m) and limited sampling rates (10 Hz), approximating the characteristics of commercial millimeter-wave radars. The model's high performance under these noisy conditions suggests a degree of robustness suitable for deployment in hardware-constrained environments.

Despite these promising results, several limitations warrant future investigation. Firstly, while the simulation mimics basic sensor noise, it does not fully replicate complex real-world factors such as multi-path interference, occlusions in multi-lane scenarios, or signal attenuation caused by adverse weather (e.g., rain or fog). Secondly, the current framework assumes homogeneous freeway traffic; its applicability to urban environments with complex interactions (e.g., intersections) remains to be tested. Thirdly, relying on a single sensor type may limit reliability.

Future work will focus on validating the model with physical sensor data, extending the framework to urban scenarios, and exploring multi-source data integration (e.g., fusing radar with camera or GPS data) to further enhance system robustness.

5 Conclusion

In this paper, we propose a novel FCM-RF-SMOTE framework for traffic state classification, integrating Fuzzy C-Means (FCM) clustering, Random Forest (RF) classification, and the Synthetic Minority Over-sampling Technique (SMOTE). The primary contributions of this study are as follows.

1. A standardized classification system was derived based on the clustering centers identified by the FCM algorithm. The analysis defined the following data-driven thresholds:

smooth (>110 km/h), stable (80–110 km/h), congested (40–80 km/h), and severely congested (<40 km/h).

2. A novel FCM-RF-SMOTE framework is proposed to effectively address data imbalance, improving the minority class (severe congestion) proportion from 3.67% to 19.83%, and enhancing classification accuracy from 77.67% to 97.80%.
3. The fuzzy clustering feature of FCM quantifies ambiguous traffic state boundaries, improving upon traditional threshold-based methods and better reflecting the continuity of traffic flow.
4. The framework is initially validated via SUMO simulation with simplified radar-like noise ($\sigma = 0.1$ m, 10 Hz). Its performance in real-world scenarios requires further field testing.

Looking ahead, future work should focus on three main directions: (1) Long-Term Deployment, evaluating the framework's performance in real-world traffic management systems over extended periods and under diverse conditions; (2) Integration with Smart City Infrastructure, such as connected vehicles and intelligent traffic signals, to enable more dynamic and adaptive traffic management; and (3) Advanced Algorithms for Predictive Analysis, exploring the use of advanced algorithms like Convolutional Neural Networks (CNNs) and Transformers to further enhance the framework's ability to predict traffic congestion and optimize traffic flow in real time. In conclusion, the FCM-RF-SMOTE framework provides a robust and efficient approach to traffic state classification, with significant implications for intelligent transportation systems and traffic management.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RC: Methodology, Validation, Conceptualization, Writing – review and editing. AL: Writing – review and editing, Conceptualization. XS: Writing – original draft, Investigation. FL: Writing – original draft, Software. NL: Writing – original draft, Formal Analysis. YW: Project administration, Writing – original draft, Supervision. LY: Validation, Writing – original draft. QY: Writing – original draft, Visualization, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgements

Song Shujian was involved in submitting the manuscript and replying to reviewers on behalf of the corresponding author.

Conflict of interest

Author RC was employed by Henan Zhongyuan High-speed Zhengluo Construction Co., Ltd. Author AL was employed by Jiangxi Communications Investment Group Co., Ltd. Authors XS, FL, NL, YW, and LY were employed by CCCC Highway Consultants Co., Ltd.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

References

- Ahmed, S. R., and Cook, A. R. (1982). Application of time-series analysis techniques to highway incident detection. *Transp. Res. Rec.* 841, 19–21.
- Author Anonymous (2012). *Interim technical requirements for road network operation monitoring and service*. Beijing: China Communications Press.
- Barth, M., and Boriboonsomsin, K. (2008). Real-world carbon dioxide impacts of traffic congestion. *Transp. Res. Rec.* 2058, 163–171. doi:10.3141/2058-20
- Bauza, R., Gozalvez, J., and Sanchez-Soriano, J. (2010). Road traffic congestion detection through cooperative vehicle-to-vehicle communications[C]//*IEEE local Computer Network Conference*. IEEE, 606–612.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16 (1), 321–357. doi:10.1613/jair.953
- Collins, J. F., Hopkins, C. M., and Martin, J. A. (1979). Automatic incident Detection-TRRL algorithms HIOCC and PATREG. *TRRL Suppl. Rep.* 526. *Transp. Road Res. Laboratory*.
- Cook, A. R., and Cleveland, D. E. (1974). Detection of highway capacity-reducing incidents by traffic-stream measurements. *Transp. Res. Rec.* 495, 1–11.
- Dudek, C. L., Messer, C. J., and Nuckles, N. B. (1974). Incident detection on urban highways. *Transp. Res. Rec.* 495, 12–24.
- Hawas, Y. E. (2007). A fuzzy-based system for incident detection in urban street networks. *Transp. Res. Part C Emerg. Technol.* 15 (2), 69–95. doi:10.1016/j.trc.2007.02.001
- Hsiao, C. H., Lin, C. T., and Cassidy, M. (1994). Application of fuzzy logic and neural networks to automatically detect highway traffic incidents. *J. Transp. Eng.* 120 (5), 753–772. doi:10.1061/(asce)0733-947x(1994)120:5(753)
- Jiang, J., Chen, Q., Xue, J., Wang, H., and Chen, Z. (2020). A novel method about the representation and discrimination of traffic State. *Sensors* 20, 5039. doi:10.3390/s20185039
- Jin, X., Ma, W. F., Zhong, R. X., and Jiang, G. G. (2023). An efficient variational Bayesian algorithm for calibrating fundamental diagrams and its probabilistic sensitivity analysis. *Transp. B Transp. Dyn.* 11 (1), 1616–1641. doi:10.1080/21680566.2023.2231159
- Li, S., and Yoon, H.-S. (2023). Sensor fusion-based vehicle detection and tracking using a single camera and radar at a traffic intersection. *Sensors* 23 (10), 4888. doi:10.3390/s23104888
- Liu, Q., Lu, J., and Chen, S. (2014). Design and analysis of traffic incident detection based on random forest. *J. Southeast Univ. Engl. Ed.* 1, 88–95. doi:10.3969/j.issn.1003-7985.2014.01.017
- Manual, H. C. (2010). *HCM2010[M]. Transportation research board*. Washington, DC: National Research Council.
- Martin, P. T., Perrin, J., Hansen, B., Kump, R., and Moore, D. (2001). *Incident detection algorithm evaluation*. Salt Lake City, UT, United States: Utah Department of Transportation.
- Ming, W., Qian, W., Shan, W. L., Guo, J., Li, W., Lin, W., et al. (2023). Taxi drivers' traffic violations detection using random forest algorithm: a case study in China. *Traffic Inj. Prev.* 24 (1/4), 362–370. doi:10.1080/15389588.2023.2191286
- Nanthawichit, C., Nakatsuji, T., and Suzuki, H. (2003). Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a highway. *Transp. Res. Rec. J.* 1855, 49–59. doi:10.3141/1855-06
- Park, H., and Yoon, S. H. (2024). Deep reinforcement learning for base station switching scheme with federated LSTM-based traffic predictions. *ETRI J.* 46 (3), 379–391. doi:10.4218/etrij.2023-0065
- Park, H.-C., Kim, D.-K., and Kho, S.-Y. (2018). Bayesian network for highway traffic State prediction. *Transp. Res. Rec.* 2672, 124–135. doi:10.1177/0361198118786824
- Payne, H. J., and Tignor, S. C. (1978). Highway incident-detection algorithms based on decision trees with states. *Transp. Res. Rec.* (682).
- Puangnak, K., and Chivapreecha, S. (2019). Comparative study of threshold selection for incident detection based on California Algorithm. 911, 914. doi:10.1109/ECTI-CON47248.2019.8955226
- Ranpura, P., Gujar, R., and Shukla, V. (2024). Estimation of traffic delay due to traffic control elements using Bayesian Optimized Predictive Model for heterogeneous traffic conditions. *10th Int. Conf. Control, Decis. Inf. Technol. (CoDIT).0*. doi:10.1109/CoDIT62066.2024.10708404
- Reddy, D. P., Manjunath, M., Rohith, M., Reddy, N. M., and Satyanarayana, A. (2024). Deep CNN model for condition monitoring of road traffic: an application of computer vision. *Turkish J. Comput. Math. Educ. (TURCOMAT)* 14, 1362–1370. doi:10.61841/turcomat.v14i03.14525
- Shaaban, K., Davoodi, S. R., and Shaaban, K. (2024). Analyzing autonomous vehicle collision types to support sustainable transportation systems: a machine learning and Association rules approach. *Sustainability* 16, 9893. doi:10.3390/su16229893
- Shang, Q., Feng, L., and Gao, S. (2021). A hybrid method for traffic incident detection using random forest-recursive feature elimination and long short-term memory network with bayesian Optimization Algorithm. *Qual. Control, Trans.* 9 (1), 1219–1232. doi:10.1109/access.2020.3047340
- Sharma, O., Dash, S., and Sial, M. R. (2023). "Road traffic congestion detection through cooperative vehicle-to-vehicle communications[C]," in *2023 4th IEEE global conference for advancement in technology (GCAT)*. IEEE, 1–6.
- Sheu, J. B., and Ritchie, S. G. (1998). A new methodology for incident detection and characterization on surface streets. *Transp. Res. Part C Emerg. Technol.* 6 (5–6), 315–335. doi:10.1016/s0968-090x(99)00002-9
- Sommer, C., Joerer, S., Segata, M., Tonguz, O., Cigno, R. L., and Dressler, F. (2011). How shadowing hurts vehicular communications and how dynamic beaconing can help. *IEEE INFOCOM*, 110–115. doi:10.1109/INFCOM.2013.6566745
- Stiller, C., Puente León, F., and Kruse, M. (2025). Information fusion for automotive applications – an overview. *Inf. Fusion* 12 (4), 244–252. doi:10.1016/j.inffus.2011.03.005
- Wang, Q. (2019). *Research on method for traffic status identification of highway basic Sections[D]*. Nanjing: Southeast University.
- Wang, B., Sun, J., Wang, W., Xu, Z., Tian, T., Wang, Y., et al. (2018). "Real time detection of traffic signal running State and remote alarm for fault information at road intersection," in *Proceedings of the 2018 24th international conference on automation and computing* (Newcastle upon Tyne, UK: ICAC), 478–482.
- Yu, Q., Feng, Z., Xu, H., and Guangli, R. (2015). Study on the recognition model of highway traffic State. *J. Transp. Eng. Inf. No. 2 Vo 1*.
- Yuan, Y. (2020). "Application of intelligent technology in urban traffic congestion," in *Proceedings of the 2020 international conference on computer engineering and application (ICCEA)* (Guangzhou, China).
- Zahid, M., Chen, Y., Jamal, A., and Memon, M. Q. (2020). Short term traffic state prediction via hyperparameter optimization based classifiers. *Sensors* 20, 685. doi:10.3390/s20030685
- Zhao, P., Geng, D., She, S., and Duan, M. (2024). Road traffic safety status analysis and prediction based on dynamic bayesian network. *J. Phys. Conf. Ser.* 2868, 012028. doi:10.1088/1742-6596/2868/1/012028