

Article

Signal Phase-Driven Traffic State Characterization: A Precision Approach for Urban Intersection Analysis

Zhengjun Li ¹, Feng Luo ¹, Liangjie Xu ^{2,*} , Xinquan Zu ², Yichen Xu ² and Feng Ji ² 

¹ CCCC Second Highway Consultants Co., Ltd., Wuhan 430056, China; xinling5xinyu@163.com (Z.L.); lfengxx@126.com (F.L.)

² School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan 430063, China; zuxq2001@163.com (X.Z.); 347003@whut.edu.cn (Y.X.); finn1025@163.com (F.J.)

* Correspondence: laurrie119@163.com

Abstract

Data-driven traffic state identification classifies traffic states by mining the operational characteristics of traffic flow, so as to support traffic control and management. Due to the influence of signal phases, it is difficult to accurately classify nonlinear traffic flow. To address this issue, this study proposes a refined traffic state identification method based on intersection phases, aiming at the complex traffic flow characteristics of signalized intersections. First, the collected traffic flow data are segmented and reorganized according to the signal phase scheme. Training samples with “traffic state” labels are obtained using the Fuzzy C-Means (FCM) clustering algorithm optimized by the Rime Optimization Algorithm (RIME), i.e., the RIME-FCM algorithm. Then, the XGBoost model optimized by Bayesian optimization (BO-XGBoost-SHAP) is applied to achieve accurate identification of traffic data. Finally, the effectiveness of the proposed method is verified using real road data. The results show that this method can effectively estimate the traffic flow state of signalized intersections and avoid “false congestion”.

Keywords: BO-XGBoost; ITS; RIME-FCM; SHAP; signalized; traffic state recognition

1. Introduction

With the growth of travel demand, traffic issues such as congestion and accidents have become increasingly prominent. These issues seriously affect economic development and people’s quality of life, and accurately obtaining road traffic states is the most important and fundamental task for addressing these traffic problems. For instance, traffic control and management systems require accurate and timely road traffic state data to formulate corresponding traffic control strategies. Additionally, the quality of traffic services depends on the quality of collected road traffic state information.

Existing traffic state identification methods can be categorized into two types: model-driven methods and data-driven methods. Model-driven methods establish traffic flow models based on traffic flow theory, such as the Macroscopic Fundamental Diagram (MFD) [1–5], the Lighthill–Whitham–Richards (LWR) model [6–11], and the Kalman Filter (KF) model [12–15]. However, such model-driven methods usually rely on idealized assumptions and struggle to fully capture the complexity of real-world traffic. This study focuses on data-driven traffic state identification at signalized intersections.

Data-driven methods do not require explicit functional relationships. They can mine the correlations between traffic flow operational characteristics and traffic states from



Academic Editor: Ján Dižo

Received: 31 May 2026

Revised: 27 July 2026

Accepted: 28 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

massive historical traffic flow data, and then identify the current traffic state based on real-time traffic flow data. Common methods include BIC [16], NN [17–19], SVM [20–22], and KNN [23,24]. For example, Xia et al. [16] selected traffic volume, vehicle speed, and vehicle occupancy as traffic state evaluation indicators, used BIC to fit the statistical characteristics of traffic flow data, and estimated traffic states by combining the rate of divergence change. With the rise of deep learning algorithms, some studies have used traffic flow images in surveillance data to estimate traffic states [25,26]. These studies indirectly infer traffic states by estimating vehicle speeds from traffic surveillance images. For instance, Sreekumar et al. [26] adopted a background enhancement algorithm to extract moving vehicle speeds and used the optical flow method to estimate vehicle speeds.

With the in-depth research on deep learning, some studies have applied deep learning algorithms to traffic state identification [27–31]. Tang et al. [28] used the area ratio of vehicles to roads in traffic images to quantify traffic state levels. They proposed a new traffic state identification framework named TrafficNet, which integrates the visual attention mechanism of Squeeze-and-Excitation (SE) and deep metric learning based on Modified Euclidean Triplet Loss (MEIT). Lu et al. [31] proposed a deep integrated neural network that fuses Graph Convolutional Network (GCN), Gated Recurrent Unit (GRU), and attention mechanism. Beklaryan et al. [32] developed an intelligent transportation system simulation model with adaptive signal timing capability based on fuzzy clustering, while Prabuchandran et al. [33] were among the first to introduce the multi-agent reinforcement learning framework into traffic signal control scenarios, laying the algorithmic foundation for distributed cooperative control across multiple intersections.

Although the aforementioned studies have achieved significant progress in hybrid modeling, uncertainty handling, and deep learning applications, they generally overlook a critical physical constraint when applied to urban signalized intersections: signal phase. Additionally, when existing methods are applied to urban intersections, they almost all treat traffic flow as continuous data. This makes it impossible to distinguish between two scenarios: normal and predictable stopping queues caused by red signals, and abnormal and problematic random congestion caused by supply–demand imbalance. This misjudgment of “false congestion” as real congestion is the primary source of errors in intersection state identification, thereby misleading traffic managers.

Most existing data-driven studies treat traffic flow as continuous data and overlook the physical constraints imposed by signalized intersections. In contrast, this study proposes a signal phase-driven theoretical framework. By fundamentally separating phase-specific traffic flow data, the framework successfully distinguishes predictable queues caused by red signals from abnormal congestion resulting from supply–demand imbalance, thereby effectively eliminating the long-standing issue of “false congestion” misclassification in intersection traffic state analysis.

To achieve robust traffic state definition under highly fluctuating traffic conditions, a globally optimized RIME-FCM algorithm is introduced. By integrating the Rime Optimization Algorithm (RIME) into the traditional Fuzzy C-Means (FCM) clustering process, the proposed method avoids convergence to local optima and ensures the robustness and reliability of traffic state classification in signalized intersection environments characterized by discontinuous traffic flow patterns and ambiguous state boundaries.

To provide transparent and actionable decision support for traffic management, the SHAP framework is embedded into the BO-XGBoost classifier to overcome the “black-box” limitation of high-performance machine learning models. This methodological design not only achieves high classification accuracy but also quantifies the dynamic marginal contribution of each traffic parameter to the recognition results, offering traffic managers transparent and interpretable insights into the underlying causes of congestion.

2. Research Methodology

To address the misjudgment issue of traditional methods caused by overlooking signal phases, this study proposes a “signal phase-driven” refined identification framework—which is not a single algorithm but comprises three closely coordinated components. First, for data preprocessing, a “signal phase-driven data preprocessing framework” is adopted to fundamentally isolate traffic flow data of different phases and eliminate the interference of “false congestion”; subsequently, for state definition, an optimized “RIME-FCM” clustering model is used to conduct scientific and robust state classification on the separated phase data; and finally, for identification and interpretation, a “BO-XGBoost-SHAP” model is employed to train a high-performance classifier based on the defined state labels and provide interpretability for the identification results. This section elaborates on the specific implementation of these three components in detail.

2.1. Traffic State Classification at Intersections

The classification of traffic states at intersections is fundamental to traffic state division. To determine the optimal number of clusters, this study adopts an improved PBM_F index as an evaluation indicator for fuzzy clustering validity. The expression for this index is formulated as

$$PBM_F = \frac{(D/E) \times S}{v} \quad (1)$$

Here, D represents the total distance from the data points to the global centroid, E denotes the intra-cluster compactness, S signifies the inter-cluster separation, and v is the number of clusters. By calculating the PBM_F index for different numbers of clusters (as shown in Figure 1), the results indicate that the index reaches its maximum value when the number of clusters is four. Consequently, the traffic state is categorized into four classes: Free-Flow, Moderately Free-Flow, Congested, Severely Congested. The selection of discriminant parameters, including speed, traffic volume, and space occupancy, is based on the “Evaluation Method for Road Traffic Congestion” (GA/T 115-2020 [34]).

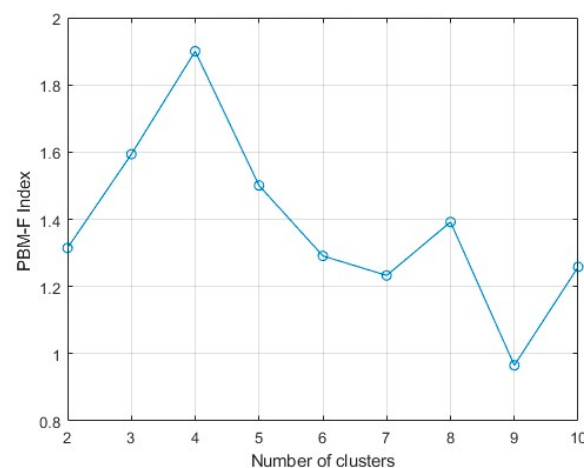


Figure 1. PBM_F index for different numbers of clusters.

Traditional Fuzzy C-Means (FCM) clustering algorithms are highly sensitive to initial cluster centers, which can lead to slow convergence and instability. To address these limitations, this study introduces the RIME-FCM algorithm, which optimizes the initial cluster centers using the Rime Optimization Algorithm (RIME) algorithm. RIME mimics frost ice

growth, incorporating soft frost search strategies and hard frost piercing mechanisms to update particle positions. The fitness function for this algorithm is defined as follows:

$$f = \frac{1}{J_{FCM}(U, C)} \quad (2)$$

Algorithm Implementation:

1. Initialize parameters: fuzzy index m , population size N , maximum iterations $T_{\max}$, termination threshold ε , etc.
2. Update cluster centers C and membership matrix U .
3. Define the fitness function and iteratively update particle positions via RIME.
4. Update the global best particle. If the maximum iteration count is reached, proceed to Step 5; otherwise, return to Step 2.
5. Terminate the algorithm and use the particle with the highest fitness as the initial cluster center for FCM to partition samples.

Studies demonstrate that the RIME-FCM algorithm enhances the quality of initial cluster centers, outperforming traditional FCM in convergence speed and stability.

2.2. Traffic State Recognition Model

XGBoost is an enhanced algorithm based on the traditional Gradient Boosting Decision Tree (GBDT) algorithm. To prevent overfitting, XGBoost introduces additional decision trees for feature splitting, where each new tree is designed to fit the residuals of the previous model's predictions. It features a customized objective function that incorporates regularization to prevent overfitting and is designed with parallel processing capabilities. The primary distinction between XGBoost and GBDT lies in the definition of the objective function.

The objective function of XGBoost is defined as follows:

$$J^t = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + c \quad (3)$$

In the equation, y_i denotes the true data, $\hat{y}_i^{(t-1)} + f_t(x_i)$ represents the sum of the predicted values from the first $t - 1$ learners and the t -th learner, l is the model's loss function, $\Omega(f_t)$ refers to the regularization strategy, and c is a constant.

The regularization term is defined as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (4)$$

In the formula, T denotes the number of leaf nodes, γ is the regularization term for the number of leaves, λ is the regularization term for the leaf weights, and ω represents the leaf weight values.

XGBoost primarily enhances the loss function with regularization and uses a reductive approach to prevent overfitting. It also applies Taylor expansion to the objective function, using the derived expression as a splitting criterion for tree construction. This allows the algorithm to fit the objective function while avoiding overfitting, improving recognition accuracy and generalization ability. Moreover, XGBoost supports parallel processing, which enhances computational efficiency. However, its complexity in application stems from having numerous parameters that are sensitive to tuning. To optimize these parameters effectively, this study employs Bayesian optimization, a sequential model-based optimization (SMBO) algorithm. Bayesian optimization assumes that the model trained on sample data follows a function distribution dependent on the input samples. It uses the function f

derived from previous data to determine the next sampling point and assumes f follows a Gaussian process (GP). Each input to the GP is a data variable with a Gaussian distribution, and the calculation involves determining the expectation and covariance function distribution at that point. The posterior expectation of the model is computed using the likelihood model of function f and the prior probability model of f . The assumption is as follows:

$$y = f(x) + \epsilon, \epsilon \sim N(0, \delta^2) \quad (5)$$

For the prior distribution of the model, the Gaussian process (GP) is employed to describe function f , which is determined by its mean function $m(x)$ and covariance function $k(x, x')$.

Given the data $X = \{x_1, \dots, x_n\}$, the multivariate Gaussian distribution for the distribution function $F = \{f(x_1), \dots, f(x_n)\}$ is given by

$$F \sim N(m(X), k) \quad (6)$$

In the equation, $m(x) = [m(x_1), \dots, m(x_n)]^T$ is an $n \times n$ matrix of covariance elements. The Gaussian process (GP) function returns the mean and variance of the Gaussian distribution for the predicted values of f at x .

The Bayesian optimization-based XGBoost model (BO-XGBoost) is constructed to optimize the hyperparameters of the XGBoost model for traffic state recognition. The BO algorithm is used to optimize the hyperparameters of the XGBoost model, including the number of decision trees, the minimum number of samples for node splitting, the maximum number of features for splitting, the maximum depth of the tree, and the minimum weight of the child node. The goal is to find a suitable combination of hyperparameters to enhance the model's generalization ability and improve its prediction accuracy and stability. Initially, BO randomly selects a set of hyperparameters to train a baseline XGBoost model and evaluates its performance using cross-validation. Subsequently, BO employs Bayesian principles to combine the current model performance with the prior distribution of hyperparameters to compute a potential candidate hyperparameter set. This candidate set is used to train a new XGBoost model and re-evaluate its performance. The iterative process continues, updating the posterior distribution of hyperparameters based on new observations, gradually focusing on the optimal performance region. This process repeats until a preset number of iterations or convergence criteria are met, enabling BO to efficiently search the vast hyperparameter space and significantly enhance the model's predictive power.

The basic workflow of the BO-XGBoost model is as follows, and the specific process is shown in Figure 2:

1. Initialization: Define the parameter space, randomly sample t sets of parameters, train the XGBoost model, compute the objective function values, and select the initial optimal parameters.
2. Modeling: Construct a Gaussian process regression model based on the initial parameters and their objective function values.
3. Iterative Optimization: a. Predict the mean and variance of new parameter combinations using the Gaussian process, evaluate them with an acquisition function, and select candidate parameters. b. Train the model using the candidate parameters, assess its performance on the validation set, and update the current optimal parameters. c. Add the new parameters and their objective function values to the dataset and update the Gaussian process model.
4. Termination: Repeat Step 3 until the maximum number of iterations is reached, and output the globally optimal parameter combination.

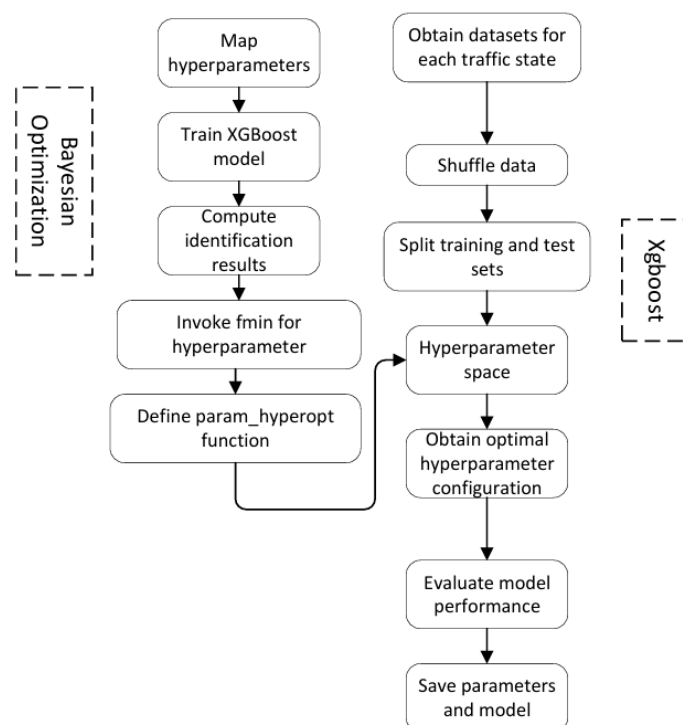


Figure 2. A refined phase-based traffic state identification method for intersections.

3. Case Study and Validation

3.1. RIME-FCM-Based State Classification

Experimental data were collected from the intersection of Jiangcheng Avenue and Taizi North Road in Wuhan. Traffic data were collected at the study intersection over a two-week observation period during three peak time windows (08:00–10:00, 12:00–14:00, and 18:00–20:00). All traffic parameters were aggregated into 1 s intervals for subsequent analysis, as shown in Figure 3. The acquired data fields include date, video time, seconds, space occupancy, traffic volume, and speed. Taking the south entrance as an example, cluster analysis was conducted separately for each lane.

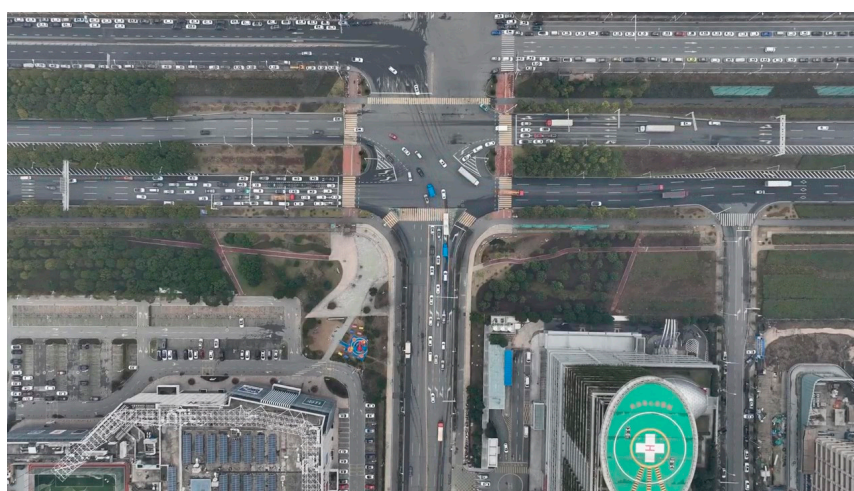


Figure 3. Satellite imagery of real-world intersections.

The parameter settings of the RIME-FCM model are presented in Table 1. Clustering centers for the four traffic states under different phases are as follows:

$$\begin{aligned}
 O_{\text{North-South Straight Phase}} &= [o_1, o_2, o_3, o_4]^T = \begin{bmatrix} 44.5560 & 21.1955 & 0.1084 \\ 34.7794 & 27.0683 & 0.1386 \\ 23.6222 & 40.3469 & 0.2037 \\ 13.1432 & 64.9329 & 0.3372 \end{bmatrix} \\
 O_{\text{North-South Left-Turn Phase}} &= [o_1, o_2, o_3, o_4]^T = \begin{bmatrix} 40.8635 & 15.0854 & 0.0768 \\ 31.7584 & 21.5739 & 0.1111 \\ 34.6167 & 31.9853 & 0.1674 \\ 18.7583 & 49.1977 & 0.2552 \end{bmatrix} \\
 O_{\text{Other Phases}} &= [o_1, o_2, o_3, o_4]^T = \begin{bmatrix} 42.7572 & 11.1442 & 0.0600 \\ 33.6017 & 11.2447 & 0.0621 \\ 34.7800 & 19.1617 & 0.1065 \\ 28.0423 & 21.2852 & 0.1190 \end{bmatrix}
 \end{aligned}$$

$[o_1, o_2, o_3, o_4]$ represent the centers of the four traffic states (free-flow, moderately free-flow, congested and severely congested) respectively. The first column of the matrix denotes speed, the second column denotes traffic volume, and the third column denotes space occupancy.

Table 1. RIME-FCM parameters.

Parameter	Value
Number of clusters v	4
Fuzzy index m	2
Termination threshold ε	1×10^{-5}
FCM maximum iterations $T_{\max}$	100
Population size n	50
Step function segments w	5
RIME maximum iterations T	30
Particle bounds $[U_{bij}, L_{bij}]$	$[0, 1]$

The data from these cluster centers provide a scientific explanation for these four states based on traffic flow theory:

State o_1 (Free-Flow): Exhibits the highest speed, low volume, and the lowest occupancy. This aligns with the free-flow state in traffic flow theory, where vehicles operate almost unimpeded.

State o_2 (Moderately Free-Flow): Shows medium–high speed, medium volume, and low–medium occupancy. This corresponds to a stable or moderately free-flow state, where traffic density is increasing, but operation remains stable.

State o_3 (Congested): Speed drops significantly, while volume and occupancy both rise markedly. This indicates the traffic flow is at or near its critical state, definable as the congested state.

State o_4 (Severely Congested): Presents the lowest speed, high volume and the highest occupancy. This clearly points to a severely congested or forced-flow state, characterized by stop-and-go traffic and low network efficiency.

Furthermore, comparing the cluster centers of different phases (e.g., straight vs. left-turn) reveals significant differences in their traffic characteristics (such as free-flow speed), which in turn proves the necessity of the “signal phase-driven” preprocessing adopted in this study.

(a) Convergence Speed Comparison

As shown in Table 2, RIME-FCM achieves optimal clustering in 13 iterations, whereas traditional FCM requires 52 iterations. RIME-FCM exhibits superior convergence speed.

Table 2. Objective function values of FCM and RIME-FCM.

Iteration	FCM Objective Value	RIME-FCM Objective Value
1	50,998.63148	51,494.03848
2	39,002.25526	38,577.73595
.....		
12	6582.68822	6513.849579
13	6581.851655	6513.769915
.....		
51	6513.76991	
52	6513.76991	

(b) Stability Comparison

Running both algorithms 10 times (Table 3), RIME-FCM consistently achieves identical fitness values, whereas FCM shows higher variability. RIME-FCM demonstrates enhanced stability and faster convergence.

Table 3. Performance comparison between FCM and RIME-FCM.

Run	RIME-FCM			FCM		
	RIME-FCM Initial Fitness	RIME-FCM Final Fitness	Iteration	FCM Initial Fitness	FCM Final Fitness	Iteration
1	51,494.03848	6513.76991	13	50,998.63148	6513.76991	31
2	50,904.21492	6513.76991	12	51,066.04527	6513.76991	28
3	50,940.2544	6513.76991	12	51,456.47817	6513.76991	30
4	50,757.03484	6513.76991	12	51,257.78629	6513.76991	27
5	51,633.80256	6513.76991	13	51,300.05247	6513.76991	29
6	51,635.71617	6513.76991	14	51,318.85041	6513.76991	28
7	51,208.9507	6513.76991	14	51,687.51411	6513.76991	29
8	51,739.75369	6513.76991	15	51,157.57602	6513.76991	31
9	51,724.71547	6513.76991	16	51,253.74389	6513.76991	28
10	51,494.03848	6513.76991	16	51,691.90501	6513.76991	27

3.2. Traffic State Recognition Using the BO-XGBoost-SHAP Model

The hyperparameter optimization of the BO-XGBoost model identified the optimal hyperparameter combination at the 120th iteration, achieving an AUC score of 0.977 on the train set, as detailed in Table 4. The optimal hyperparameter combination for the BO-XGBoost model was (452.0, 17.8, 362.0, 52.0, 1.2).

Table 4. Bayesian optimization hyperparameter comparison.

Iter	Target	n_Estimators	Min_Samples_Split	Max_Leaves	Max_Depth	Min_Child_Weight
1	0.792	3.0	6.57	55.0	12.0	1.5
2	0.821	12.0	22.97	123.0	34.0	3.3
.....						
120	0.973	452.0	17.8	362.0	52.0	1.2
121	0.955	421.0	22.6	360.0	53.0	1.6

3.2.1. Model Performance Evaluation

The traffic state dataset comprises three parameters: traffic flow, space occupancy, and average speed. To ensure the precision of the traffic state recognition model, the

experimental data were randomly shuffled, with the first 80% used as the training set and the remaining 20% as the test set. The model was compared against the baseline XGBoost, Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN) algorithms, and the classification accuracy is presented in Table 5. The results show that all five traffic state recognition models achieved over 90% accuracy, with the BO-XGBoost model attaining an average accuracy of 98.90%, outperforming the baseline XGBoost, SVM, RF, and KNN in both accuracy and stability.

Table 5. Comparison of model recognition accuracy.

Iterations	Bo-XGBoost	XGBoost	SVM	RF	KNN
1	99.87%	99.87%	96.85%	98.74%	98.43%
2	99.87%	99.87%	96.33%	96.89%	98.56%
3	92.74%	99.74%	96.19%	97.2%	98.56%
4	90.00%	80.00%	96.06%	98.24%	98.81%
5	99.87%	99.87%	95.27%	92.86%	98.29%
6	89.87%	99.87%	96.98%	99.74%	97.24%
7	100.00%	100.00%	97.11%	98.73%	98.95%
8	100.00%	100.00%	95.67%	99.87%	97.64%
9	99.87%	99.87%	97.24%	97.2%	98.82%
.....					
Average Accuracy	98.90%	92.90%	93.35%	91.83%	90.43%

The classification accuracy and runtime performance of each model are presented in Table 5 and Table 6, respectively. As indicated in Table 6, BO-XGBoost exhibited the shortest average runtime of 3.4429 s, comparable to KNN and shorter than RF. In summary, BO-XGBoost demonstrated superior performance in both accuracy and efficiency compared to the baseline XGBoost, SVM, RF, and KNN, highlighting its effectiveness in traffic state recognition.

Table 6. Model runtime comparison.

Iterations	BO-XGBoost	XGBoost	SVM	RF	KNN
1	3.4203	3.96861	3.7893	4.0683	3.6223
2	3.8697	3.64877	3.9229	4.2164	3.4100
3	3.8481	3.30196	3.5913	3.8408	3.4030
4	3.5395	3.51944	3.6462	4.6383	3.3543
5	3.4687	3.57451	3.5126	3.7416	3.3560
6	3.5136	3.64227	3.5495	3.6792	3.5585
7	3.4587	3.46019	3.6128	4.0333	3.4358
8	3.8966	3.96904	3.6382	3.6862	3.4109
9	3.4637	3.59621	3.5286	4.0974	3.4877
.....					

3.2.2. Ablation Study

To eliminate the risk of information leakage caused by temporal autocorrelation and to further verify the model's true generalization capability, an independent, time-separated validation set was introduced. This validation dataset consists of 10 min of continuous traffic flow data collected on a separate day, entirely outside the original two-week observation period. When the BO-XGBoost model (trained on the original dataset) was applied to this independent data for traffic state recognition, it achieved a high accuracy of 96.17%.

Specifically, the 10 min validation set consists of 600 samples. The signal phase distribution includes 46.7% green-light samples and 53.3% red-light samples. Regarding

the traffic states, the distribution is: 180 samples of free-flow, 100 samples of moderately free-flow, 150 samples of congested, and 170 samples of severely congested.

To further elucidate the model’s classification performance across different traffic states, Table 7 presents the confusion matrix evaluated on this independent validation set. The matrix indicates that the BO-XGBoost model correctly identifies the vast majority of samples across all four categories. Most misclassifications occur between adjacent states, such as confusing free-flow with moderately free-flow, which aligns with the continuous nature of traffic flow transitions. Notably, the model demonstrates robust capability in distinguishing normal queuing (congested) from gridlock (severely congested).

Table 7. Confusion matrix on the independent validation set.

True State/Predicted State	Free-Flow	Moderately Free-Flow	Congested	Severely Congested
Free-Flow	175	5	0	0
Moderately Free-Flow	3	92	5	0
Congested	0	2	142	6
Severely Congested	0	0	2	168

Correspondingly, Table 8 details the class-level precision, recall, and F1-score. The model achieves an F1-score of 97.76% for the free-flow state and 97.67% for the severely congested state. The consistently high macro-averaged F1-score (95.72%) across the four classes confirms that the overall high accuracy (96.17%) is not an artifact of class imbalance, but rather a reflection of the model’s balanced and reliable identification capability across all traffic conditions.

Table 8. Class-level performance metrics of BO-XGBoost.

Traffic State	Precision	Recall	F1-Score
Free-Flow	98.31%	97.22%	97.76%
Moderately Free-Flow	92.93%	92.00%	92.46%
Congested	95.30%	94.67%	94.98%
Severely Congested	96.55%	98.82%	97.67%
Macro Average	95.77%	95.68%	95.72%

Furthermore, to substantiate the effectiveness of the proposed methodology, an ablation study was conducted in which the raw traffic flow data was processed without phase-driven segmentation. The results indicate that during the two-week observation period, 32.11% of the red-light samples were erroneously classified as the “congested” state when phase constraints were ignored. This robustly demonstrates the indispensability of phase-driven segmentation for enhancing the accuracy of traffic state estimation at signalized intersections with interrupted flow.

3.2.3. Explainability Analysis of Models Based on SHAP

SHAP is a unified framework that generates a “SHAP value” for each prediction sample. This value is the sum of the contributions allocated to each feature and is interpreted as the feature’s contribution to the prediction. It is calculated based on the marginal contribution of the feature across all possible feature combinations:

$$g(z') = \Phi_0 + \sum_{j=1}^M \Phi_j z'_j \quad (7)$$

Among them, g represents the variables, and the explanatory model $z' \in \{0, 1\}^M$ is interpreted as a joint vector. M is the maximum joint scale. $\Phi_j \in R$ denotes the feature attributes for feature j . The attributes of each variable typically depend on three factors: the number of times a feature is used to split the feature space across all trees, the split count weighted by the number of instances that pass through these splits, and the total reduction in training loss when a feature is used for splitting.

SHAP-based explanations allow us to examine the important statistical features in traffic state recognition and provide insights into improving the recognition process. The model's predictions are interpreted at both local and global levels. For each traffic flow parameter, the contribution of each statistical feature to the recognition result is visualized through local SHAP values. Figure 4 illustrates the interpretation of the traffic flow state for the 1021st and 2101st samples in the dataset. In the figure, features that increase recognition are represented by red bars, while those that decrease it are shown in blue. The base value represents the model's baseline prediction, which is the average expected output across the entire dataset. To deeply illustrate how feature importance dynamically adapts to signal phases, we analyze two distinct scenarios. For the 1021st sample, recorded during a red-light phase, the traffic state is recognized as congested. Due to the signal interruption, the "speed" feature drops significantly, contributing the most to the model's estimate with a strong positive impact (pushing the prediction towards a higher congestion severity). Conversely, for the 2101st sample, recorded during a green-light phase, the XGBoost model estimates the traffic flow as moderately free-flow. In this context, the high "speed" feature again acts as the most influential variable, but it exerts a significant negative impact. Additionally, the global impact of these features on the model's predictions is analyzed based on the trends in their SHAP values. Figure 5 presents the distribution of SHAP values for statistical features sorted by their absolute SHAP values, clearly indicating the influence of each feature. In Figure 5, data samples are represented by colored points ranging from blue to red, ordered by the original values of the statistical features. Overall, the "speed" feature has the most significant impact on state recognition. As previously noted, the substantial differences in traffic parameters across different phases greatly affect traffic flow state recognition, and a rough classification of traffic states can have a negative impact on recognition. Applying SHAP to analyze the feature importance of each traffic parameter provides key insights into the components of traffic state recognition.

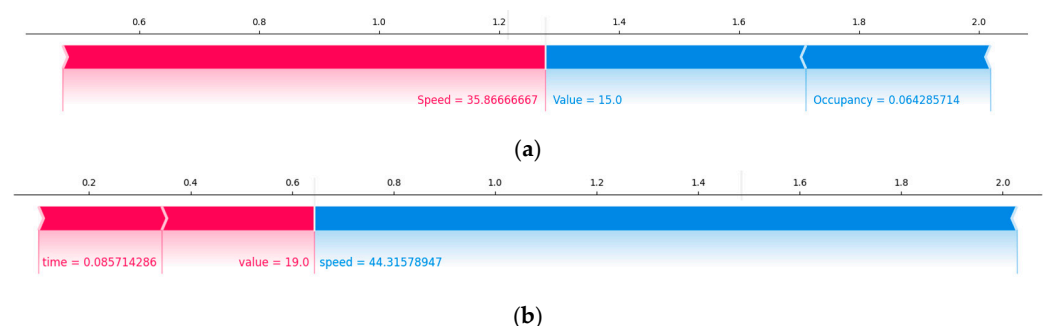


Figure 4. Local explanation of traffic flow state identification. (a) SHAP value of sample 1021. (b) SHAP value of sample 2101.

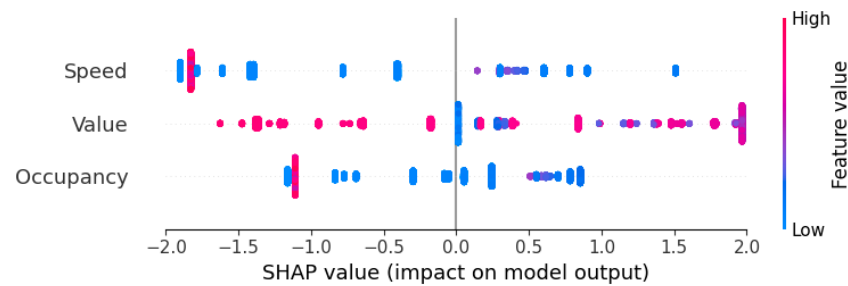


Figure 5. Global explanation of traffic flow state identification.

4. Conclusions

To address the complex traffic flow characteristics of urban intersections, this study proposes a refined traffic state identification method based on intersection phases. By integrating the Fuzzy C-Means (FCM) clustering algorithm optimized by the Rime Optimization Algorithm (RIME-FCM) and the XGBoost model optimized by Bayesian optimization (BO-XGBoost-SHAP), this method systematically resolves the limitations of traditional traffic state identification methods in intersection scenarios.

First, it proposes a signal phase-driven data preprocessing framework for traffic flow data at intersections. Raw data are segmented and reorganized according to the signal phase scheme, which eliminates mutual interference between traffic flows of different phases. On this basis, a modified PBM_F index is introduced to determine the optimal number of clusters. Second, it designs the RIME-FCM algorithm, which optimizes initial cluster centers by simulating the rime growth process, significantly improving the convergence speed and stability of the traditional FCM algorithm. Finally, it constructs the BO-XGBoost model, which dynamically optimizes hyperparameter combinations via Gaussian process regression. Meanwhile, the SHAP framework is incorporated to enable interpretability analysis of the model, quantifying the contribution of each feature to traffic state identification.

The refined traffic state identification method proposed in this study fills the research gap in current traffic state identification methods for signalized intersections. It can provide data support for active traffic control strategies.

Although the proposed model achieves high accuracy in traffic state classification, several limitations remain. First, the training and testing datasets were divided using a random shuffling strategy. Given the strong temporal autocorrelation of traffic flow data, this approach may introduce information leakage between adjacent samples. As a result, the reported performance may be somewhat optimistic and may not fully reflect the model's generalization capability under unseen traffic conditions or future time periods.

Author Contributions: Conceptualization, Z.L.; methodology, Z.L. and X.Z.; software, Z.L. and F.L.; validation, F.L. and X.Z.; formal analysis, X.Z.; investigation, F.L. and F.J.; resources, F.J.; data curation, Z.L., Y.X. and F.J.; writing—original draft preparation, Z.L. and F.J.; writing—review and editing, L.X.; visualization, Y.X.; supervision, L.X.; project administration, L.X.; funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by the Hubei Provincial Department of Science and Technology through the Hubei Provincial Key Research and Development Program titled “Research and Application of Cross-Road Domain Traffic Flow State Identification and Transfer Technology Driven by Spatiotemporal Data” under Grant 2023BAB022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Authors Zhengjun Li and Feng Luo were employed by the company CCCC Second Highway Consultants Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Dervisoglu, G.; Gomes, G.; Kwon, J.; Horowitz, R.; Varaiya, P. Automatic calibration of the fundamental diagram and empirical observations on capacity. In *Transportation Research Board 88th Annual Meeting*; TRB: Washington, DC, USA, 2009; Volume 15, pp. 31–59.
2. Zhong, R.; Chen, C.; Chow, A.; Pan, T.; Yuan, F.; He, Z. Automatic calibration of fundamental diagram for first-order macroscopic freeway traffic models. *J. Adv. Transp.* **2016**, *50*, 363–385. [\[CrossRef\]](#)
3. Herrera, J.; Work, D.; Herring, R.; Ban, X.; Jacobson, Q.; Bayen, A. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transp. Res. Part C-Emerg. Technol.* **2010**, *18*, 568–583. [\[CrossRef\]](#)
4. Seo, T.; Kusakabe, T.; Asakura, Y. Traffic State Estimation with the Advanced Probe Vehicles using Data Assimilation. In *Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Gran Canaria, Spain, 15–18 September 2015; pp. 824–830. [\[CrossRef\]](#)
5. Seo, T.; Kusakabe, T.; Asakura, Y. Calibration of Fundamental Diagram Using Trajectories of Probe Vehicles: Basic Formulation and Heuristic Algorithm. *Transp. Res. Procedia* **2017**, *21*, 6–17. [\[CrossRef\]](#)
6. Haj-Salem, H.; Lebacque, J. Reconstruction of false and missing data with first-order traffic flow model. *Transp. Res. Rec. J. Transp. Res. Board* **2002**, *1802*, 155–165. [\[CrossRef\]](#)
7. Claudel, C.; Bayen, A. Guaranteed bounds for traffic flow parameters estimation using mixed Lagrangian-Eulerian sensing. In *Proceedings of the 2008 46th Annual Allerton Conference on Communication, Control, And Computing*, Monticello, IL, USA, 23–26 September 2008; Volumes 1–3, p. 636. [\[CrossRef\]](#)
8. Xia, C.; Cochrane, C.; DeGuire, J.; Fan, G.; Holmes, E.; McGuirl, M.; Murphy, P.; Palmer, J.; Carter, P.; Slivinski, L.; et al. Assimilating Eulerian and Lagrangian data in traffic-flow models. *Physica D* **2017**, *346*, 59–72. [\[CrossRef\]](#)
9. Whitham, G.B. *Linear and Nonlinear Waves*; Wiley-Interscience: New York, NY, USA, 1974.
10. Mihaylova, L.; Boel, R. A particle filter for freeway traffic estimation. In *Proceedings of the 2004 43rd IEEE Conference on Decision and Control (CDC)*, Nassau, Bahamas, 14–17 December 2004; Volume 1–5, pp. 2106–2111. [\[CrossRef\]](#)
11. Yuan, Y.; Duret, A.; van Lint, H. Mesoscopic traffic state estimation based on a variational formulation of the LWR model in Lagrangian-space coordinates and Kalman filter. *Transp. Res. Procedia* **2015**, *10*, 82–92. [\[CrossRef\]](#)
12. Wang, Y.; Papageorgiou, M. Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. *Transp. Res. Part B-Methodol.* **2005**, *39*, 141–167. [\[CrossRef\]](#)
13. Zhen, F.; Jabari, S.; Liu, H.; Lin, D. Traffic state estimation using stochastic Lagrangian dynamics. *Transp. Res. Part B-Methodol.* **2018**, *115*, 143–165. [\[CrossRef\]](#)
14. Ouessai, A.; Keche, M. Real-time freeway traffic state estimation based on the second-order divided difference Kalman filter. *Transp. Telecommun. J.* **2018**, *20*, 114–122.
15. Nantes, A.; Ngoduy, D.; Bhaskar, A.; Miska, M.; Chung, E. Real-time traffic state estimation in urban corridors from heterogeneous data. *Transp. Res. Part C-Emerg. Technol.* **2016**, *66*, 99–118. [\[CrossRef\]](#)
16. Xia, J.; Huang, W.; Guo, J. A clustering approach to online freeway traffic state identification using ITS data. *KSCE J. Civ. Eng.* **2012**, *16*, 426–432. [\[CrossRef\]](#)
17. Dougherty, M.; Cobbett, M. Short-term inter-urban traffic forecasts using neural networks. *Int. J. Forecast.* **1997**, *13*, 21–31. [\[CrossRef\]](#)
18. Jiang, X.; Adeli, H. Dynamic wavelet neural network model for traffic flow forecasting. *J. Transp. Eng.* **2005**, *131*, 771–779. [\[CrossRef\]](#)
19. Thakuriah, P.; Tilahun, N. Incorporating Weather Information into Real-Time Speed Estimates: Comparison of Alternative Models. *J. Transp. Eng.* **2013**, *139*, 379–389. [\[CrossRef\]](#)
20. Deng, C.; Wang, F.; Shi, H.; Tan, G. Real-time freeway traffic state estimation based on cluster analysis and multiclass support vector machine. In *Proceedings of the International Workshop on Intelligent Systems and Applications*, Wuhan, China, 23–24 May 2009.
21. Siddique, C.; Ban, X.J. Traffic state estimation based on vehicle trajectory segmentation. In *Proceedings of the Transportation Research Board 96th Annual Meeting Transportation Research Board*, Washington, DC, USA, 8–12 January 2017.

22. Cao, J.; Fang, Z.; Qu, G.; Sun, H.; Zhang, D. An accurate traffic classification model based on support vector machines. *Int. J. Netw. Manag.* **2017**, *27*, e1962. [[CrossRef](#)]
23. Oh, S.; Byon, Y.J.; Yeo, H. Improvement of search strategy with k-nearest neighbors approach for traffic state prediction. *IEEE Trans. Intell. Transp. Syst.* **2015**, *17*, 1146–1156. [[CrossRef](#)]
24. Xu, D.; Wang, Y.; Peng, P.; Beilun, S.; Deng, Z.; Guo, H. Real-time road traffic state prediction based on kernel-KNN. *Transp. A-Transp. Sci.* **2020**, *16*, 104–118. [[CrossRef](#)]
25. Faro, A.; Giordano, D.; Spampinato, C. Evaluation of the traffic parameters in a metropolitan area by fusing visual perceptions and CNN processing of webcam images. *IEEE Trans. Neural Netw.* **2008**, *19*, 1108–1129. [[CrossRef](#)] [[PubMed](#)]
26. Sreekumar, U.K.; Devaraj, R.; Li, Q.; Liu, K. TPCAM: Real-time traffic pattern collection and analysis model based on deep learning. In Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 4–8 August 2017; pp. 1–4.
27. Reddy, S.; Ram, B.; O’Byrne, M.; Vanajakshi, L.; Ghosh, B. Alternative approach to traffic state analysis on Indian roads using image processing. *Proc. Inst. Civ. Eng.-Transp.* **2019**, *172*, 336–346. [[CrossRef](#)]
28. Tang, F.; Fu, X.; Cai, M.; Lu, Y.; Zeng, Y.; Zhong, S.; Huang, Y.; Lu, C. Multilevel Traffic State Detection in Traffic Surveillance System Using a Deep Residual Squeeze-and-Excitation Network and an Improved Triplet Loss. *IEEE Access* **2020**, *8*, 114460–114474. [[CrossRef](#)]
29. Liu, Y.; Lyu, C.; Zhang, Y.; Liu, Z.; Yu, W.; Qu, X. DeepTSP: Deep traffic state prediction model based on large-scale empirical data. *Commun. Transp. Res.* **2021**, *1*, 100012. [[CrossRef](#)]
30. Zhang, K.; Feng, X.; Jia, N.; Zhao, L.; He, Z. TSR-GAN: Generative Adversarial Networks for Traffic State Reconstruction with Time Space Diagrams. *Phys. A-Stat. Mech. Its Appl.* **2022**, *591*, 126788. [[CrossRef](#)]
31. Lu, W.; Yi, Z.; Wu, R.; Rui, Y.; Ran, B. Traffic speed forecasting for urban roads: A deep ensemble neural network model. *Phys. A-Stat. Mech. Its Appl.* **2022**, *593*, 126988. [[CrossRef](#)]
32. Beklaryan, A.L.; Beklaryan, L.A.; Akopov, A.S. Simulation model of an intelligent transportation system for the “smart city” with adaptive control of traffic lights based on fuzzy clustering. *Bus. Inform.* **2023**, *17*, 70–86. [[CrossRef](#)]
33. Prabuchandran, K.J.; Hemanth Kumar, A.N.; Bhatnagar, S. Multi-agent reinforcement learning for traffic signal control. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 2529–2534. [[CrossRef](#)]
34. GA/T 115-2020; Evaluation Methods for Road Traffic Congestion Levels. Ministry of Public Security of the People’s Republic of China: Beijing, China, 2020.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.