



A novel traffic accident detection method with comprehensive traffic flow features extraction

Liping Zhu^{1,2} · Bingyao Wang^{1,2} · Yihan Yan³ · Shuang Guo⁴ · Gangyi Tian^{1,2}

Received: 24 May 2021 / Revised: 11 February 2022 / Accepted: 31 March 2022 / Published online: 28 April 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

With the rapidly increasing of automobiles, traffic accidents are gradually becoming more frequent. This creates a great need for effective traffic anomaly detection algorithms. Existing methods shed light on directly inferring the abnormalities from traffic flow, which is short in features extraction and representation of traffic flows. In this paper, we propose three new traffic flow features, namely the road congestion, the traffic intensity, and the traffic state instability, for more comprehensive traffic status representation and anomaly detection. Residual analysis, quadratic discrimination, multi-resolution wavelet analysis are integrated for the extraction of the aforementioned features, which will be applied for the downstream tasks of traffic anomaly detection. Experimental results reveal that accident identification based on the proposed features is more effective than the raw traffic flow, which is supposed to provide an alternative approach for further applications and studies.

Keywords Traffic accident detection · Machine learning · Feature extraction · Traffic flow features

1 Introduction

In modern transport systems, various detectors provide actionable information in critical situations, which enable us to automatically discover the abnormality of the traffic stream in time. However, due to the complexity and variability in mass traffic behavior, it is difficult to directly identify abnormal traffic events from raw observed flow measures. Therefore, more sophisticated approaches that can extract features with clean meaning and effective representation are necessary for the automatic analysis of traffic flow data.

The detection of abnormal traffic accidents has already been realized based on machine learning algorithms, artificial intelligence [27] and deep learning [10]-related technologies. Xia [28] proposed an unsupervised method based

on the sparse topic model to capture motion patterns and detect anomalies in traffic surveillance. Elizabeth Hou [12] addressed the problem of detecting anomalous activity in traffic networks where the network is not directly observed based on the Bayesian hierarchical model. Ronald D. Hagan [9] presented a case study on the analysis of New York City taxi traffic using the compound analytics framework. Silva Nuno [24] used PCA to analyze the attributes complexity of traffic features. Cuadra-Sanchez [4] focused on longitudinal traffic analysis, namely detecting sudden peak changes. Takahiro Kudo [14] detected traffic anomalies for every period of measured traffic via PCA. Youcef Djenouri [6] reviewed the use of outlier detection approaches in urban traffic analysis. Rupam Deb [5] presented a correlation-based imputation method to improve the quality of traffic flow. Shu-Bin Li [15] realized accident detection by taking into account the traffic ratio at the entrances and crossways. Seyed Hessam-Allah [11] provided a novel rule-based method to predict traffic accident severity according to user's preferences.

Recently, Zheng Zhao [31] discussed a novel traffic forecast model based on long short-term memory (LSTM) network. Meanwhile, Mehrannia and BagiSiamese, et al [18] also investigate the deep representation of loop detector data using LSTM for automatic detection of freeway accidents. To deal with scenarios where only small datasets are avail-

✉ Bingyao Wang
2019215868@student.cup.edu.cn

¹ Beijing Key Laboratory of Petroleum Data Mining, Beijing, China

² College of Information Science and Engineering, China University of Petroleum(Beijing), Beijing, China

³ School of Applied Economics, Renmin University of China, Beijing, China

⁴ Faculty of Petroleum, China University of Petroleum (Beijing) at Karamay, Xinjiang, China

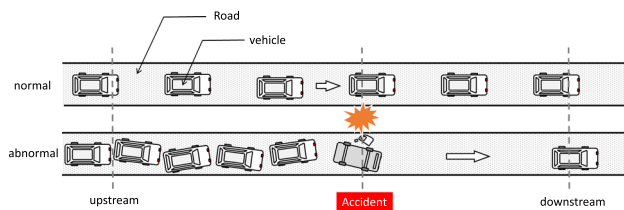


Fig. 1 A visual statement of the vehicle collision accident, the normal condition (top) vs abnormal condition (bottom) after accident

able for training, Sabour and Rao, et al. [22] further develop the Siamese neural network-based DeepFlow to automatically analyze traffic flow data. Meanwhile, XGBoost [21], ensemble support vector machine [26], isolation forest [17], and other machine learning algorithms [23] are also applied for flow data-based abnormal traffic status detection.

In this study, we focus on the feature extraction of traffic flow data for more effective abnormal traffic events identification. Here the term *abnormal traffic event* refers to an exception point in time that the traffic system behaves abnormally and is significantly different from the previous normal behavior. This can be caused by natural factors (heavy rainfall in short-terms for example), or human factors such as traffic accidents. Fig. 1 shows the abnormal scene of vehicle collision. Since traffic flow parameters often show a significant trend under normal circumstances, the abnormal vehicle collision will cause a significant impact on traffic parameters [7] [20].

However, the occurrence of traffic accidents is real-time, complex, and sporadic. Meanwhile, relationships between these accidents and their reflection on raw traffic flow parameters are hard to be summarized as clear rules. This brings great difficulties to automatic anomaly detection of traffic status. Based on the I80-E highway traffic flow data and accident records provided by the US PeMS system [30], we propose three new traffic flow features, namely the road congestion, the traffic intensity, and the traffic state instability, for more comprehensive traffic status representation and anomaly detection. Our study is executed according to the flowchart shown in Fig. 2.

2 Methodology

2.1 Road congestion features

Road congestion refers to the traffic phenomenon caused by the traffic vehicle surge. McMaster algorithm is an algorithm based on the theoretical model of highway traffic [29] state catastrophe. However, during our application of the McMaster algorithm to the I80-E highway data, we find there are three limitations in the McMaster algorithm: (1) The critical

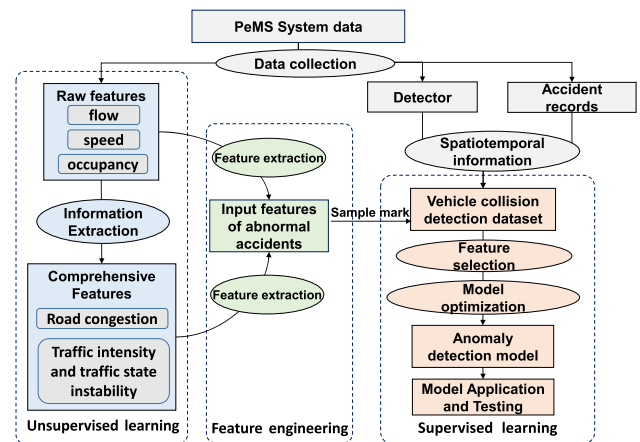


Fig. 2 Flowchart of traffic accident detection with the proposed feature extraction method

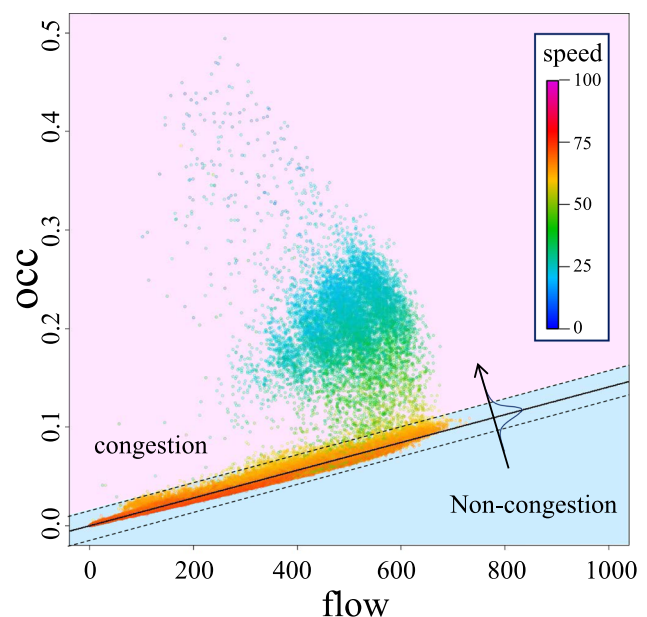


Fig. 3 Scatter plot of the traffic flow data

values of occupancy and volume greatly depend on human analysis. (2) The parameter and configuration of different traffic detector/detectors are diverse, which needs us to analyze case-by-case. (3) The McMaster algorithm ignores the observed traffic parameters of speed. Meanwhile, it lacks consideration of other kinds of anomalies.

Therefore, we develop a new method to eliminate the above limitations. Fig. 3 demonstrates the traffic flow data of the detector S400430, where the X-axis represents the flow, Y-axis is the occupancy, and the color illustrates the speed. When the speed is faster than 50, the road is in the non-congestion state, and the flow is linear related to the occupancy. On the contrary, in the state of road congestion (speed is lower than 50), the flow and occupancy no longer

satisfy the simple linear relationship, which forms several clusters away from the linear normal trends.

According to the above characteristics of road congestion, we design road congestion identification features based on the Quadratic Discrimination Analysis (QDA) algorithm [3]. We utilize the following linear basis function model to map the flow X_{flow} into the occupancy X_{occ} for the non-congestion state:

$$X_{occ} = a + bX_{flow} \quad (1)$$

where the constant a represents the vertical intercept, and b represents the slope of the linear basis function. Then, the residual E_{occ} between the predicted occupancy X_{occ} and the ground truth X_{occ}^* is adopted to measure the state of from the:

$$E_{occ} = X_{occ}^* - X_{occ}. \quad (2)$$

Then, the non-congestion occupancy of each point can be estimated by the linear relationship, while these observations are roughly labeled by the 3σ criterion [25] according to the residual difference with the real occupancy. Substituting each point flow X_{flow} into the above linear relationship, it can be found that the E_{occ} basically satisfies the Gaussian distribution with a mean of 0. Since the residual of the data points in the uncongested state is basically in the normal range, the residual value in the congestion state is obviously large. According to the 3σ criterion, the point of $E_{occ} > 3\sigma$ can be considered to be in a road congestion state, where σ is the standard deviation of E_{occ} . Finally, the Quadratic Discrimination Analysis (QDA) algorithm is used for supervised learning to predict congestion probability. The conclusion based on the 3σ criterion is only a rough discriminant conclusion, whether it is crowded or not. In order to obtain more accurate conclusions, the QDA algorithm [3] is used to obtain a classification prediction model. Applying the model to the detector data, the congestion probability of each point can be achieved, indicating the degree of road congestion.

2.2 Traffic intensity features

Traffic intensity refers to the number of vehicles detected by a detector per unit time. Since traffic patterns for working days and holidays are different, therefore, traffic data of Saturdays, Sundays, and legal holidays in the United States are marked for further analysis. Then, the short-term historical data of flow data are collected for the forecast of trend values in each day. The median of non-holiday traffic intensity in the first two weeks (about 10 days) is used as the trend values for working days. Meanwhile, the median of holidays traffic

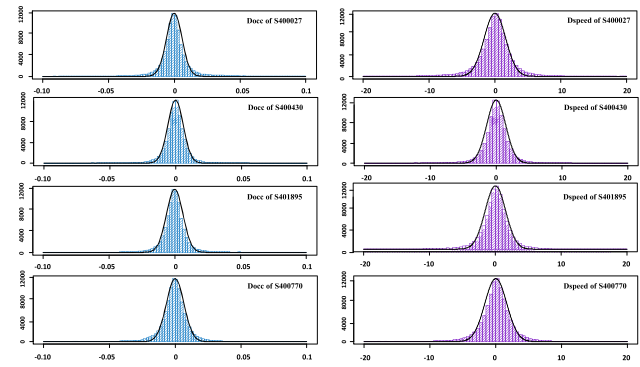


Fig. 4 Histogram of D_{occ} and D_{speed} for partial detectors

intensity in the first four weeks (about 8 days) is used as the trend values for holidays.

We take the following approach for the extraction of traffic intensity anomaly features. Firstly, the differences between actual and predicted occupancy and speed are calculated for each moment. Assuming that X_{occ} and X_{speed} represent actual data values of occupancy and speed at a certain time, respectively, while T_{occ} and T_{speed} indicate their corresponding predicted values. When there is no anomaly, D_{occ} and D_{speed} tend to 0. When abnormalities occur and traffic increases, D_{occ} will be far greater than 0, and D_{speed} will be far less than 0. In contrast, D_{occ} will be far less than 0, and D_{speed} will be far greater than 0 if traffic decreases.

Then, the cumulative probability density functions of D_{speed} and D_{occ} according to normal distribution are estimated. As can be observed in Fig. 4, we find that the empirical distribution of D_{occ} and D_{speed} of each detector is quite similar with the probability density curves of corresponding normal distribution. To further validate this observation, we performed the Kolmogorov–Smirnov normal likelihood test [16] for D_{occ} and D_{speed} , while the testing results are presented in Table 1.

It is confirmed that the distribution of D_{occ} and D_{speed} on each detector obey the normal distribution (with expectation 0 and statistic variances, respectively) with large confidence values. Therefore, the abnormal traffic intensity features P_{more} and P_{less} can be calculated from the cumulative probability function of approximated normal distributions. According to the definition of cumulative probability function of normal distribution, the cumulative probability density functions of D_{occ} and D_{speed} are:

$$P_{occ}(D_{occ}) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{D_{occ}}{\sqrt{2}\delta_{occ}} \right) \right) \quad (3)$$

$$P_{speed}(D_{speed}) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{D_{speed}}{\sqrt{2}\delta_{speed}} \right) \right) \quad (4)$$

Table 1 Kolmogorov–Smirnov normality test result of D_{occ} and D_{speed} for partial detectors

Detector	The normal confidence coefficient of D_{occ}	The normal confidence coefficient of D_{speed}
S400027	0.656	0.683
S400430	0.772	0.705
S401895	0.719	0.715
S400770	0.703	0.658
S401195	0.801	0.794
S401892	0.693	0.644
S408632	0.614	0.696
S401269	0.678	0.712

where δ_{occ}^2 and δ_{speed}^2 are the statistic variances of D_{occ} and D_{speed} , respectively. Then, the abnormal traffic intensity features P_{more} and P_{less} are defined as:

$$P_{anorm} = P_{occ}(D_{occ}) * (1 - P_{speed}(D_{speed})) \quad (5)$$

$$P_{more} = \begin{cases} 2.0 * (P_{anorm} - 0.5); & P_{anorm} > 0.5 \\ 0.0; & P_{anorm} \leq 0.5 \end{cases} \quad (6)$$

$$P_{less} = \begin{cases} 2.0 * (0.5 - P_{anorm}); & P_{anorm} \leq 0.5 \\ 0.0; & P_{anorm} > 0.5 \end{cases} \quad (7)$$

The value ranges of P_{more} and P_{less} are both (0,1). A high P_{more} value indicates an abnormal increase in traffic intensity, while a higher P_{less} indicates an abnormal decrease. Compared with directly using D_{occ} and D_{speed} for the abnormal traffic intensity representation, P_{more} and P_{less} effectively eliminate the magnitude differences of D_{occ} and D_{speed} among different detectors.

2.3 Traffic state instability features

We also develop a wavelet analysis-based approach to extract features for the representation of traffic state instability. This approach is applied to the series of three raw observed traffic parameters (flow, speed, and occupancy), and the corresponding local activity and fluctuation intensity features are obtained for the representation of overall and local variations of traffic flows.

The traffic state instability features are extracted based on the multi-resolution wavelet analysis framework. It is necessary to eliminate local fluctuations in the flow data that are independent of the overall trend changing. Therefore, we calculate the overall trend of traffic flow through a discrete binary wavelet transform [19]-based frequency domain smoothing algorithm. In our approach, the Daubechies

wavelet basis is applied. According to Daubechies wavelet function $\psi(t)$ and scale function $\phi(t)$:

$$\begin{cases} \psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j}t - k) \\ \phi_{j,k}(t) = 2^{-j/2} \phi(2^{-j}t - k) \end{cases} \quad (8)$$

Then, the trend $f_a^j(t)$ and detail $f_d^j(t)$ on the j -th scale can be constructed step by step:

$$\begin{cases} f_a^j(t) = \sum_k c_{j,k} \phi_{j,k}(t); & c_{j,k} = \int f(t) \overline{\phi_{j,k}(t)} dt \\ f_d^j(t) = \sum_k d_{j,k} \psi_{j,k}(t); & d_{j,k} = \int f(t) \overline{\psi_{j,k}(t)} dt \end{cases} \quad (9)$$

In the above definition, $c_{j,k}$ is the scale expansion coefficient and $d_{j,k}$ is the wavelet expansion coefficient. Setting J as an arbitrary scale, traffic flow data can be reconstructed by:

$$f(t) = \sum_j \sum_k d_{j,k} \psi_{j,k}(t) + \sum_k c_{J,k} \phi_{J,k}(t). \quad (10)$$

The multi-resolution wavelet algorithm can be applied for band-pass filtering of traffic flows. We will decompose the raw signal $f(t)$ with J -level multi-resolution wavelet transform at first, while scale expansion coefficients $c_{j,k}$ and wavelet expansion coefficients $d_{j,k}$ at all levels are obtained. By thresholding $c_{j,k}$ and $d_{j,k}$, the corresponding coefficients of components outside the band-pass region are assigned with 0. Then, the modified coefficients $\overline{c_{j,k}}$ and $\overline{d_{j,k}}$ are used for wavelet reconstruction to obtain the overall trends of traffic flows. According to the above multi-resolution wavelet filtering method, 8-level Daubechies wavelet transform is performed on the road occupancy data of detectors from low to high, while the high-frequency part of each level is eliminated step by step. We retain 4 to 8 wavelets to smooth the series of flow, speed, and occupancy, which means the time-domain resolution of estimated overall trends is about half an hour.

Finally, we calculate the mean absolute difference (fluctuation intensity) between traffic flow data and overall trend within a given range to indicate the local fluctuation. Traffic flow data are obtained from PeMS System every 5 minutes. At the current time point, 13 points (about one hour) are taken forward and backward, respectively, to calculate the local activity and fluctuation intensity. Assuming that the value of traffic flow data at time t is X_t and the corresponding overall trend value is T_t , the formula for calculating local activity A_t and fluctuation intensity F_t at that time is as follows:

$$A_t = \frac{1}{13} \sum_{i=t-6}^{t+6} \left(T_i - \frac{1}{13} \sum_{i=t-6}^{t+6} T_i \right)^2 \quad (11)$$

$$F_t = \frac{1}{13} \sum_{i=t-6}^{t+6} |X_i - T_i|. \quad (12)$$

According to the above formulas, six values can be calculated from three raw observed traffic parameters. We name them with the terms of *local activity* and *fluctuation intensity* for flow, speed, and occupancy, respectively.

3 Experiment

3.1 Dataset collection and feature extraction

Our experiments are based on the traffic data of the I80-E highway in 2016 from the US PeMS system [30]. The spatiotemporal features (location, date, and time) of traffic detectors and their three kinds of series data, namely speed, flow, and occupancy are collected for feature extraction. 1524 traffic records with categories *Collision Enrt*, *Collision Minor Injuries*, *Collision No Injuries*, *Collision Unknown Injuries*, *Hit and Run With Injuries*, and *Hit and Run No Injuries* are selected to form the positive sample set for vehicle collision accidents. Meanwhile, 5000 negative samples are also randomly collected from normal time without any traffic event records (the workflow for negative sample collection is presented in Fig. 5). These 6524 samples form the experimental dataset for our following study. Then, we use the proposed methods (refer to Sect. 3) to extract 60 input features. 48 upstream and downstream state features are extracted at first. These traffic accident identification features are extracted from series data of upstream and downstream detectors, within 15 minutes before and after the vehicle collision records. They are all basic mean statistics of the raw traffic flows and the extracted feature flows. The meanings and definitions of the 48-dimensional upstream and downstream features are listed in Table 2. Then, 12 California

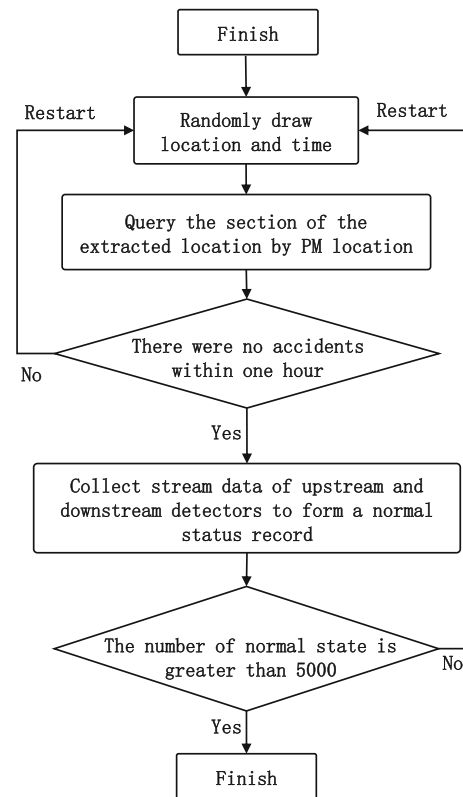


Fig. 5 Flowchart for negative sample collection workflow

algorithm features are also extracted referring to the California algorithm [13]:

$$OCCDF = OCC_{S1,t} - OCC_{S2,t} \quad (13)$$

$$OCCRDF = \frac{OCC_{S1,t} - OCC_{S2,t}}{OCC_{S1,t}} \quad (14)$$

$$DOCCTD = \frac{OCC_{S2,t-2} - OCC_{S2,t}}{OCC_{S2,t-2}}. \quad (15)$$

In the above definition, S1 represents the upstream detector and S2 is the downstream detector. We simply extend the California algorithm to other traffic flow parameters (not only the occupancy flow data), while names and definitions of these features are presented in Table 3.

3.2 Vehicle collision accidents detection modeling

Based on the collected 6524 samples and 60 input features, we view the detection of vehicle collision accidents as a binary classification task, and learn classification models with machine learning algorithms for further accidents detection:

1) Feature Analysis & Selection: To overcome the “dimension disaster” problem, the importance of 60 features is evaluated by the MDA (Mean Decrease in Accuracy)

Table 2 Features based on the mean of upstream and downstream states

Traffic flow data	A	B	C	D
flow	FLOW.MBF1	FLOW.MAF1	FLOW.MBF2	FLOW.MAF2
occupancy	OCC.MBF1	OCC.MAF1	OCC.MBF2	OCC.MAF2
speed	SPEED.MBF1	SPEED.MAF1	SPEED.MBF2	SPEED.MAF2
P_{more}	MORE.MBF1	MORE.MAF1	MORE.MBF2	MORE.MAF2
P_{less}	LESS.MBF1	LESS.MAF1	LESS.MBF2	LESS.MAF2
congestion	SATU.MBF1	SATU.MAF1	SATU.MBF2	SATU.MAF2
flt.flow	FLT.FLOW.MBF1	FLT.FLOW.MAF1	FLT.FLOW.MBF2	FLT.FLOW.MAF2
flt.occ	FLT.OCC.MBF1	FLT.OCC.MAF1	FLT.OCC.MBF2	FLT.OCC.MAF2
flt.speed	FLT.SPEED.MBF1	FLT.SPEED.MAF1	FLT.SPEED.MBF2	FLT.SPEED.MAF2
act.flow	ACT.FLOW.MBF1	ACT.FLOW.MAF1	ACT.FLOW.MBF2	ACT.FLOW.MAF2
act.occ	ACT.OCC.MBF1	ACT.OCC.MAF1	ACT.OCC.MBF2	ACT.OCC.MAF2
act.speed	ACT.SPEED.MBF1	ACT.SPEED.MAF1	ACT.SPEED.MBF2	ACT.SPEED.MAF2

A: Mean value of the upstream detector before accidents occur. B: Mean value of the upstream detector after accidents occur. C: Mean value of the downstream detector before accidents occur. D: Mean value of the downstream detector after accidents occur.

Table 3 Features based on generalized California algorithm

Traffic flow data	R1	R2	R3
flow	FLOW.CLF1	FLOW.CLF2	FLOW.CLF3
occupancy	OCC.CLF1	OCC.CLF2	OCC.CLF3
speed	SPEED.CLF1	SPEED.CLF2	SPEED.CLF3

R1: California algorithm features referred to Eq (13). R2: California algorithm features referred to Eq (14). R3: California algorithm features referred to Eq (15)

method of random forest [8]. Its core idea is to investigate the influence of the random disturbance of each feature on the prediction accuracy [1], which is evaluated on the OOB (Out Of Bagging) test sets through ensemble learning. Through our analysis, 7 features are reserved for further classification modeling, namely SATU.MAF1, SATU.MBF1, MORE.MAF1, F -LT.OCC.MAF1, LESS.MAF2, MORE.MAF2, and LE -SS.MBF1. In addition to these filtered features, the spatial and temporal information (represented by the hour and related detector positions) are included in the inputs.

2) Algorithm Parameters Tuning: Six classification algorithms including the *linear discriminant analysis* (LDA), *quadratic discriminant analysis* (QDA), *neural network* (NNET), *classification support vector machine* (CSVM), *classification and regression trees* (CART), and *classification random forest* (CRF) are applied. These algorithms take the above 9 dimensional features as inputs, while directly output whether a vehicle collision accident happens for this moment. According to the normal range of each algorithm, we use grid search [2] to evaluate the accuracy of model generalization under various parameter combinations and then determine the optimal parameter combinations. The best algorithm parameter combination and its generalization accuracy evaluation are shown in Table 4.

3) Algorithm Evaluation & Selection: We use 20-fold cross-validation to validate the generalization performance of established models. The evaluated results are shown in Table 5, while the best performance is highlighted with bold text. It is easy to see that the accuracy of the CRF model is 0.911, which is significantly higher than other models. The prediction accuracy and precision of CRF are both above 0.85. Because of the serious imbalance of sample data, the recall rate of CRF is only 0.655. However, it is also significantly higher than other models. Therefore, the CRF algorithm is finally selected for modeling. Then, the final vehicle collision detection model is constructed with all 6234 samples.

4 Application and comparison

We further applied the proposed method to traffic data of I80-E highway in 2017. Table 6 shows the performance evaluation results of various baseline models [4,17,22,22,31], while the better performances in each pair are highlighted with bold texts. CRF is short for the supervised random forest approach we used in the previous section, SPC represents the sudden peak change-based method proposed by Cuadra-Sanchez [4], iForest is the isolation forest-based unsupervised abnormal analysis approach, which has been applied in [17] and [22], LSTM is the supervised LSTM-based method [31],

Table 4 The result of hyper-parameter tuning

algorithm	Optimal parameter	Accuracy estimate
NNET	size = 1 , decay = 0.1	0.888
CSVM	cost = 3.0, sigma = 0.2	0.886
CART	maxdepth = 15, minsplit = 20	0.901
CRF	ntry= 7, minsplit = 3	0.910

The best results are highlighted in bold

Table 5 Performances of Candidate Algorithms

Algorithm	Accuracy	Precision	Recall Rate
LDA	0.858	0.876	0.452
QDA	0.873	0.894	0.540
NNET	0.889	0.888	0.496
CSVM	0.885	0.883	0.472
CART	0.904	0.912	0.618
CRF	0.911	0.920	0.655

The best results are highlighted in bold

Table 6 Detection Performances of Compared Algorithms

Algorithm	Accuracy	Precision	Recall Rate
SPC _{RTS}	0.611	0.638	0.654
SPC _{RAW}	0.604	0.622	0.602
iForest _{RTS}	0.694	0.702	0.644
iForest _{RAW}	0.601	0.607	0.617
CRF _{RTS}	0.851	0.857	0.670
CRF _{RAW}	0.750	0.741	0.488
LSTM _{RTS}	0.901	0.874	0.657
LSTM _{RAW}	0.850	0.844	0.660
DeepFlow _{RTS}	0.914	0.886	0.691
DeepFlow _{RAW}	0.867	0.865	0.688

The best results in each pair are highlighted in bold

while the idea of DeepFlow [22] is also applied to further improve the performance of LSTM-based methods. The subscript RAW represents that these algorithms are applied to the raw observed traffic flow data, while RTS means that these algorithms are applied to the further extracted features. It is easy to see that the comprehensive features indeed improved the detection of vehicle collision accidents. Besides, the performance of the CRF_{RTS} model is even slightly better than LSTM_{RAW}. Meanwhile, extracted feature flows-based LSTM_{RTS} and DeepFlow_{RTS} are also more accurate than LSTM_{RAW} and DeepFlow_{RAW}. These phenomena further confirmed the advantages of the proposed feature extraction methods.

In fact, the feature selection approach in Sect. 3 also reveals the effectiveness of the comprehensive features. Among 7 selected features with high MDA feature impor-

tance scores, SATU.MAF1 and SATU.MBF1 are based on road congestion, MORE.MAF1, LESS.MAF2, MORE.MAF2, and LESS.MBF1 are based on traffic intensity, while FLT.OCC.MAF1 is obtained based on the fluctuation strength of occupancy. In contrast, the importance scores of raw traffic data (speed, flow, and occupancy)-based classification features (including these generic California features) are relatively low.

5 Discussion and limitation

The above application and comparison confirmed the advantage of extracted comprehensive features for abnormal traffic events identification. In this section, we will further discuss why the proposed flow features are superior to the raw traffic flow data.

(1) Due to differences in road conditions and traffic environments at their locations, as well as the sensors themselves, traffic flows measured by different detectors show significant differences in the magnitude and response patterns, which means the distribution of them is diverse. However, statistical abnormal identification assumes that all the samples involved obey the same distribution. The extracted flow features adaptively eliminate distribution diversion between detectors by statistical modeling on the flow data of each detector, respectively. For example, based on the flow patterns of each detector, the road congestion features use the QDA to automatically model and convert the raw traffic flow data into congestion probability values, which approximately obey the same Bernoulli distribution. Meanwhile, the cumulative probability functions (with parameters established from flow data of each detector, respectively) also effectively eliminate the magnitude differences of D_{occ} and D_{speed} among different detectors. Through our extraction approach, irrelevant responses due to detectors or their locations are suppressed, while features associated with abnormal traffic states are retained. This greatly reduces the difficulty and robustness of the subsequent traffic events identification.

(2) The proposed flow features effectively highlight the responses of abnormal traffic states. The commonality of extracted features is that they all reflect the deviation of current states from normal states. For example, the road congestion features automatically identify the occ-flow rela-

tionship for normal non-congestion situations, while indicating the degree of road congestion based on the deviation from the normal occ-flow relationship. Traffic intensity features take the median daily trends of working and holidays as the normal states, respectively, and measure the abnormal degree of current states by statistical residual modeling. Meanwhile, traffic state instability features separate the trend (normal status) and local details (abnormal variation) through multi-resolution wavelet analysis, which also highlights the possibly abnormal fluctuation intensity of raw flow data. Since the occurrence of traffic accidents often causes traffic flows to deviate from normal statuses, flow features that highlight abnormal responses of traffic flows will effectively improve the identification of traffic events.

However, the proposed traffic flow feature extraction method still has its limitations. The studied traffic data are observed from an ideal closed highway system without horizontal crossing. This enables us to assume that the road traffic is only affected by the upstream and downstream. Therefore, for a more complicated traffic system, especially for road systems with many intersections, the changes in traffic status can be influenced by more factors. In this case, it is not certain whether the anomalous response exhibited on the traffic flow data is caused by traffic events on this road. It can be expected that the accuracy of the proposed method, in this case, will be greatly reduced. Moreover, all the proposed feature extraction methods adopt a data-driven approach. Even though this facilitates the adaptability of extracted features on specialized traffic data, it also means that the established feature extraction and accident detection models can only be applied to the studied road. If the traffic system or traffic patterns change significantly, we need to conduct the whole workflow again for updating. For example, due to the impact of COVID-19 on the frequency of people traveling, traffic conditions on the I80-E freeway may change dramatically, while the validity of the model learned from previous data cannot be guaranteed further.

6 Conclusion

In this manuscript, we proposed three new traffic flow features, namely the road congestion, the traffic intensity, and the traffic state instability, for more comprehensive traffic status representation. Based on the I80-E highway traffic flow data provided by the US PeMS system, we illustrate the application of using the extracted traffic flow features for vehicle collision accident detection. Comparative experiments reveal that the proposed comprehensive traffic features can effectively improve the performance of abnormal traffic events identification, which is worth further application. However, the proposed traffic flow feature extraction method still has its limitations. For a more complex traffic system, whether the

proposed method still works well still needs more investigation. Meanwhile, all the proposed feature extraction methods adopt a data-driven approach. This also assumes that the traffic system or traffic patterns should not change significantly. Therefore, How to adaptively update the established models for dynamically changing traffic systems is still worth more study.

Acknowledgements This work was supported by the National Key R&D Program of China (2019YFA0708304) and National Natural Science Foundation of China (Grant No. 52074323).

References

1. Auret, L., Aldrich, C.: Empirical comparison of tree ensemble variable importance measures. *Chemom. Intell. Lab. Syst.* **105**(2), 157–170 (2011)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**(2), 281–305 (2012)
3. Bhattacharyya, S., Khasnobish, A., Chatterjee, S., Konar, A., Tibarewala, D.: Performance analysis of lda, qda and knn algorithms in left-right limb movement classification from EEG data. In: 2010 International Conference on Systems in Medicine and Biology, pp. 126–131. IEEE (2010)
4. Cuadra-Sanchez, A., Aracil, J., Ramos de Santiago, J.: Proposal of a new information theory-based technique based on traffic anomaly detection analysis. *Int. J. Parallel Emerg. Distrib. Syst.* **30**(6), 464–477 (2015)
5. Deb, R., Liew, A.W.C., Oh, E.: A correlation based imputation method for incomplete traffic accident data. In: Pacific Rim International Conference on Artificial Intelligence, pp. 905–912. Springer, Cham (2014)
6. Djenouri, Y., Belhadi, A., Lin, J.C.W., Djenouri, D., Cano, A.: A survey on urban traffic anomalies detection algorithms. *IEEE Access* **7**, 12192–12205 (2019)
7. Golze, J., Feuerhake, U., Koetsier, C., Sester, M.: Impact analysis of accidents on the traffic flow based on massive floating car data. *Int. Archiv. Photogram. Remote Sens. Sp. Inf. Sci.* **43**, 95–102 (2021)
8. Gregorutti, B., Michel, B., Saint-Pierre, P.: Correlation and variable importance in random forests. *Stat. Comput.* **27**(3), 659–678 (2017)
9. Hagan, R.D., Phillips, C.A., Langston, M.A., Rhodes, B.J.: Classification and anomaly detection in traffic patterns of New York city taxis: A case study in compound analytics. In: 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 1169–1174. IEEE (2018)
10. Hao, X., Zhang, G., Ma, S.: Deep learning. *Int. J. Sem. Comput.* **10**(03), 417–439 (2016)
11. Hashmienejad, S.H.A., Hasheminejad, S.M.H.: Traffic accident severity prediction using a novel multi-objective genetic algorithm. *Int. J. Crashworthiness* **22**(4), 425–440 (2017)
12. Hou, E., Yilmaz, Y., Hero, A.O.: Anomaly detection in partially observed traffic networks. *IEEE Trans. Signal Process.* **67**(6), 1461–1476 (2019). <https://doi.org/10.1109/TSP.2019.2892026>
13. Karim, A., Adeli, H.: Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm. *J. Transp. Eng.* **128**(1), 21–30 (2002)
14. Kudo, T., Morita, T., Matsuda, T., Takine, T.: PCA-based robust anomaly detection using periodic traffic behavior. In: 2013 IEEE International Conference on Communications Workshops (ICC), pp. 1330–1334. IEEE (2013)

15. Li, S.B., Sun, T., Cao, D.N., Zhang, L.: Incident detection method of expressway based on traffic flow simulation model. *Commun. Theor. Phys.* **71**(4), 468 (2019)
16. Marsaglia, G., Tsang, W.W., Wang, J., et al.: Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **8**(18), 1–4 (2003)
17. Matousek, M., Mohamed, E.Z., Kargl, F., Bösch, C., et al.: Detecting anomalous driving behavior using neural networks. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2229–2235. IEEE (2019)
18. Mehrannia, P., Bagi, S.S.G., Moshiri, B., Al-Basir, O.A.: Deep representation of imbalanced spatio-temporal traffic flow data for traffic accident detection. *arXiv preprint [arXiv:2108.09506](https://arxiv.org/abs/2108.09506)* (2021)
19. Percival, D.B., Walden, A.T.: *Wavelet Methods for Time Series Analysis*, vol. 4. Cambridge University Press, Cambridge (2000)
20. Po, L., Rollo, F., Bachechi, C., Corni, A.: From sensors data to urban traffic flow analysis. In: 2019 IEEE International Smart Cities Conference (ISC2), pp. 478–485. IEEE (2019)
21. Qu, Y., Lin, Z., Li, H., Zhang, X.: Feature recognition of urban road traffic accidents based on GA-XGBoost in the context of big data. *IEEE Access* **7**, 170106–170115 (2019)
22. Sabour, S., Rao, S., Ghaderi, M.: Deepflow: Abnormal traffic flow detection using Siamese networks. In: 2021 IEEE International Smart Cities Conference (ISC2), pp. 1–7. IEEE (2021)
23. Salman, O., Elhajj, I.H., Chehab, A., Kayssi, A.: A machine learning based framework for IoT device identification and abnormal traffic detection. *Trans. Emerg. Telecommun. Technol.* **33**, e3743 (2019)
24. Silva, N., Shah, V., Soares, J., Rodrigues, H.: Road anomalies detection system evaluation. *Sensors* **18**(7), 1984 (2018)
25. Venables, W.N., Ripley, B.D.: *Modern applied statistics with S-PLUS*. Springer Science & Business Media (2013)
26. Wang, Z., Chu, R., Zhang, M., Wang, X., Luan, S.: An improved selective ensemble learning method for highway traffic flow state identification. *IEEE Access* **8**, 212623–212634 (2020)
27. Welsh, R.: Defining artificial intelligence. *SMPTE Motion Imag. J.* **128**(1), 26–32 (2019)
28. Xia, L.M., Hu, X.J., Wang, J.: Anomaly detection in traffic surveillance with sparse topic model. *Journal of Central South University* (2018)
29. Yang, Z.Q.: Highway traffic accident prediction based on SVR trained by genetic algorithm. *Adv. Mater. Res.* **433**, 5886–5889 (2012)
30. Zhang, L., Qi, R.: Real-time flux and density estimation of freeway traffic with decentralized speed data. In: 2017 Chinese Automation Congress (CAC) (2017)
31. Zhao, Z., Chen, W., Wu, X., Chen, P.C., Liu, J.: Lstm network: a deep learning approach for short-term traffic forecast. *IET Intel. Transp. Syst.* **11**(2), 68–75 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.