

Article

A Study of Typical P-AEB Test Scenarios Based on Accident Data

Yajun Luo , Zhenfei Zhan *, Qing Mao and Zhenxing Yi

School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China; 622230040016@mails.cqjtu.edu.cn (Y.L.); 622230990128@mails.cqjtu.edu.cn (Q.M.); yizhenxing@mails.cqjtu.edu.cn (Z.Y.)

* Correspondence: zhenfeizhan@cqjtu.edu.cn; Tel.: +86-185-8072-2167

Abstract

A large number of vulnerable road users such as pedestrians continue to be injured or killed in road accidents every year, and active safety systems such as automatic emergency braking systems are expected to improve the situation. However, automatic emergency braking systems for pedestrians have been tested in a variety of real-world scenarios. The purpose of this paper is to obtain typical P-AEB test scenarios that can reflect the real and collision scenarios through real pedestrian–vehicle crash data. By using the k-means clustering algorithm based on local outlier detection, the intersection data and the straight-road data are clustered and analyzed separately, with five types of typical P-AEB straight-road test scenarios and seven types of typical P-AEB intersection test scenarios. By comparing with the existing test protocols, the test scenarios proposed in this paper have good coverage and authenticity, and can play a guiding role in the construction of specific P-AEB system test scenarios.

Keywords: vulnerable road users; automatic emergency braking systems; pedestrian safety; clustering algorithm; test scenario

1. Introduction

According to the National Bureau of Statistics, 273,098 road traffic accidents occurred in China in 2021, resulting in 281,447 injuries and 62,218 deaths, of which 58,585 (20.8%) were injuries and 12,323 (19.8%) were deaths among vulnerable road users (VRUs) (2024) [1]. As one of the primary vulnerable groups in traffic accidents, pedestrian safety has consistently been a significant research focus within the field of automotive safety.

With the popularization of intelligent networking technologies and automatic driving technologies, the matching rate of Advanced Driving Assistance Systems (ADASs) in mass-produced vehicles is gradually increasing. Automatic emergency braking (AEB), a key function within Advanced Driving Assistance Systems, is an intelligent safety system designed to prevent or mitigate collisions by automatically applying braking based on real-time risk assessment. Automatic emergency braking, one of the ADASs, is an intelligent safety system that prevents accidents by judging the driving conditions and applying different levels of braking to the vehicle. Intelligent safety systems in automobiles are expected to improve the current situation of pedestrian accidents. The development and deployment of AEB systems require comprehensive and thorough testing, and scenario-based virtual testing has become an important part of vehicle performance verification. The coverage and danger of the virtual evaluation scenarios are crucial to the performance verification of vehicle AEB systems, so the U.S. NHTSA [2], the European Union's E-NCAP [3], and China's C-NCAP [4] and other organizations have published and continuously improved



Academic Editor: Grzegorz Sierpiński

Received: 7 January 2026

Revised: 10 February 2026

Accepted: 13 February 2026

Published: 26 February 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the

World Electric Vehicle Association.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

some human–vehicle crash test scenarios. And various scholars have employed a range of research methods to construct multiple typical pedestrian AEB test scenarios [5,6]. However, most of the pedestrian test scenarios obtained so far are on straight roads, pay less attention to the traffic environment, and the natural environment only focuses on the daytime and nighttime, plus the scenarios are all common scenarios, lacking the intersection section, high-risk scenarios, etc. Therefore, the excavation of AEB pedestrian test scenarios still needs further research.

Mining, extracting, and constructing scenario information from known data sources has become an essential method for building a virtual test scenario. Real traffic data for scenario research primarily comes from two sources: road traffic accident databases and naturalistic driving datasets. Well-known accident databases include CIDAS and NAIS in China, STATS19 in the UK, and CRSS and GIDAS in the U.S. and Germany, respectively. Naturalistic driving data, collected via instrumented vehicles, is exemplified by projects like ApolloScape in China and NGSIM or HighD internationally [7]. And many organizations at home and abroad have actively carried out related work and achieved good results, such as AppolloScape of Baidu in China and databases such as NGSIM and HighD in foreign countries. Clustering algorithms are able to classify high-dimensional traffic data based on the data distribution characteristics, so it has become an important research method commonly used to build test scenarios based on the data, and the commonly used clustering algorithms are hierarchical clustering [8,9], k-modes [10] and so on.

While data-driven scenario extraction has become a trend, two critical gaps persist in existing research. First, many clustering-based studies select scenario features arbitrarily or based solely on kinematic relevance, without quantitatively linking them to real-world injury outcomes, thus potentially overlooking high-risk factors. Second, the distinct complexity of intersection environments is often underappreciated, with many studies treating them homogeneously or focusing predominantly on straight-road scenarios.

This paper addresses these gaps by making three key contributions. Methodologically, we propose a novel risk-informed clustering framework that integrates a Random Forest-based feature importance analysis with an enhanced k-means algorithm. This ensures that the selected clustering variables are those that most significantly influence pedestrian injury severity, thereby generating scenarios that are not only statistically representative but also inherently dangerous. Contextually, we explicitly and separately analyze straight-road and intersection accident data, recognizing their fundamentally different risk profiles and traffic dynamics. Technically, we enhance the clustering robustness by incorporating Local Outlier Factor detection and hierarchical clustering initialization to mitigate the impact of noise and arbitrary initial centroids common in accident data.

Through this framework, we extract five typical straight-road and seven typical intersection P-AEB test scenarios. These scenarios extend existing test protocols by including underrepresented yet high-risk situations, such as multi-lane crossings and complex T/Y-intersection conflicts, thereby providing a more comprehensive and realistic foundation for P-AEB system development and validation.

The ultimate goal of AEB systems is to physically avoid or mitigate collisions through automated braking. While the accuracy of perception and decision-making algorithms is paramount, the system's real-world effectiveness is fundamentally bounded by vehicle dynamics and tire-road interaction. The maximum achievable deceleration, and hence the minimum stopping distance, is directly governed by the available friction coefficient between the tires and the road surface. Factors such as pavement condition, tire wear, and temperature critically influence this parameter. Consequently, a comprehensive and authentic P-AEB test scenario library must consider not only the kinematic and geometric aspects of the conflict but also the physical constraints of the vehicle's braking capability.

This study, while focusing on extracting high-risk scenario logic from accident data, acknowledges this crucial dimension. We treat road surface condition as a key environmental variable in our clustering analysis and highlight in the discussion the necessity for future work to integrate detailed vehicle dynamics parameters for complete test scenario realism.

2. Material and Methods

2.1. Data Sources

Crash data in this paper comes from the CRSS database published by the Crash Report Sampling System (CRSS) of the National Highway Traffic Safety Administration (NHTSA, Washington, DC, USA) [11]. The CRSS database publishes road traffic crashes that occurred in all regions of the United States from 2016 to 2021, based on the principle that the crash involved at least one motor vehicle on the roadway and that the crash must have resulted in property damage or injury or fatality. The accidents include different types of motor vehicles, pedestrians, and different types of two-wheeled bicyclists. The database records traffic accidents as a sample of 120 characteristic pieces of information through specific coding rules, and the data can be downloaded and used through the official website (<https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/CRSS/> accessed on 12 February 2026).

Considering the applicability of pedestrian AEB test scenarios and real data adaptability, the following cleaning rules are used for the preliminary screening of raw data:

1. The object of the accident is a passenger car and a pedestrian on the frontal collision.
2. The accident vehicle is a vehicle involved in traffic.
3. The collision is a primary collision (secondary collisions and other collisions with pedestrians are not considered).
4. Accidents occurring on straight roads and at intersections are studied separately.

Based on the above censoring rules, 1473 straight-road number samples and 1198 intersection data samples were finally selected from the original traffic data set.

2.2. Hazard Element Selection Based on Random Forest Algorithm

Based on the pedestrian accident data from the U.S. Crash Reduction Record System (CRSS), initial scenario elements have been selected for straight-road and junction scenarios, considering the four dimensions of human, vehicle, road, and environmental factors. The impact of initial scenario elements on pedestrian injury severity was quantified through a Random Forest analysis, with the selection of scenario elements with higher risk levels as clustering variables. The k-means clustering algorithm, which is based on the detection of Local Outlier Factors, was applied to obtain the straight-road and intersection test scenarios for the pedestrian–automatic emergency braking (P-AEB) system. The technical route of the study is shown in Figure 1.

2.2.1. Scene Element Selection and Coding

Considering the big difference in the complexity of pedestrian–vehicle collision scenarios on straight roads and at intersections, this paper divides the pedestrian–vehicle collision data into straight-road collision data and intersection collision data for research. Considering the pedestrian AEB test scenario requirements and CRSS accident database record information, 12 scenario elements are selected from the four dimensions of pedestrian–vehicle–road–natural environment, including pedestrian gender, pedestrian age, pedestrian relative motion, road surface condition, road gradient, road lane lines, road curvature, traffic control, pedestrian crossings, weather conditions, lighting conditions, and accident speed as shown in Figure 2. Two types of intersections and collision locations are selected for the complexity of intersection data. As shown in Figure 3.

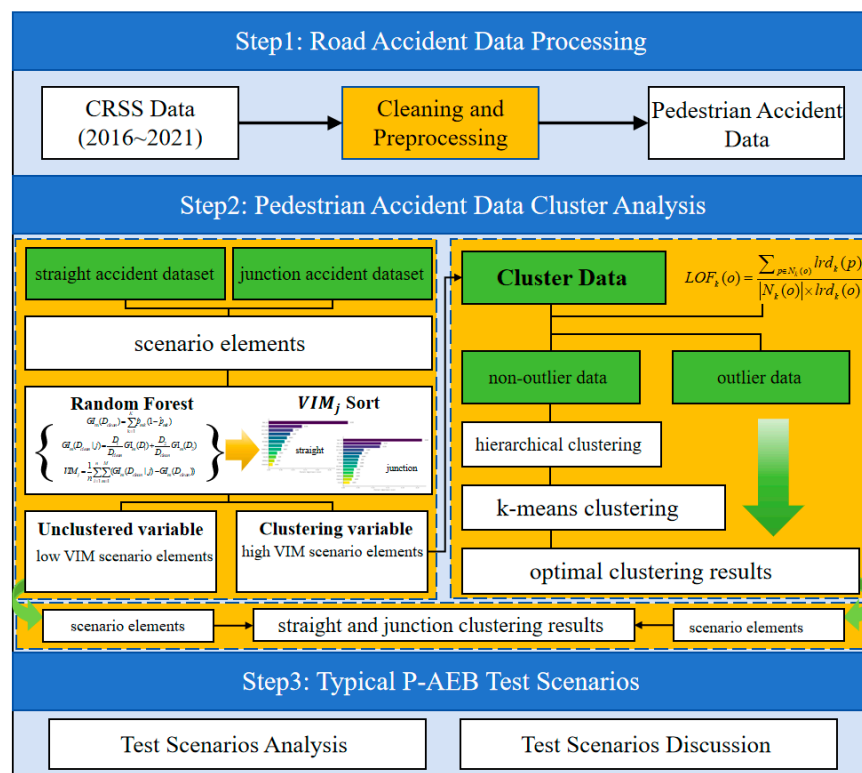


Figure 1. Technical route of this study.

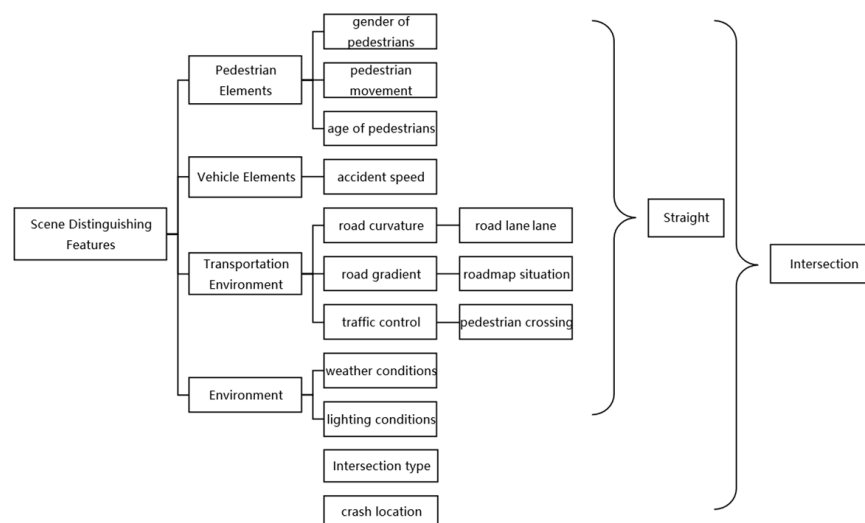


Figure 2. Selected scene characterization elements across four dimensions (pedestrian, vehicle, road, environment) for pedestrian-vehicle collision analysis.

The dataset selected through the above steps contains a total of two data formats, unordered category variables and continuous variables. The coding rules for the unordered category variables are provided by the database coding manual, and in order to facilitate the extraction of scene parameters, the various variables selected in this paper are counted and recoded.

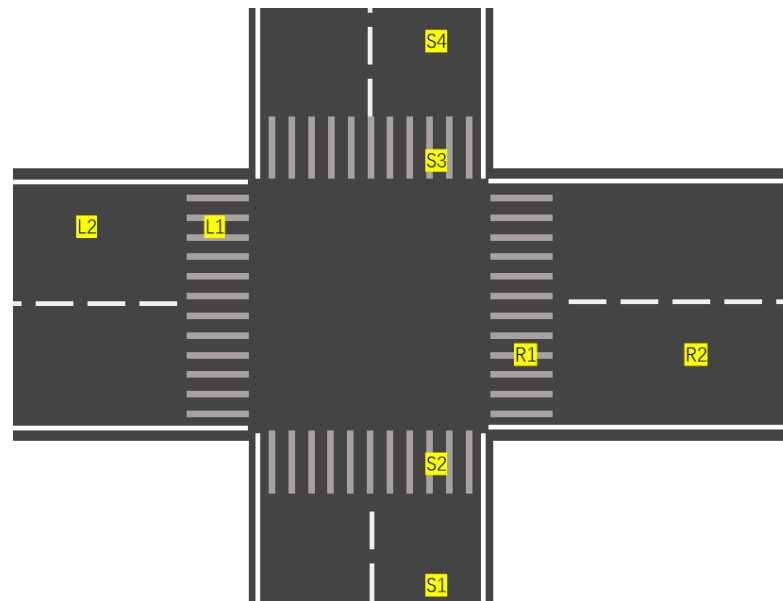


Figure 3. Classification of collision locations at intersections.

Continuous variables such as pedestrian age and vehicle speed need to be further processed; from the point of view of scene design, the specific age of pedestrians as a parameter of the scene cannot meet the test requirements of the virtual scene, because the age of pedestrians in the design of the virtual scene of the human–vehicle pre-crash information is generally reflected through the children. Therefore, this paper carries out a split-box operation on the age of pedestrians. Pedestrians under 18 years old are regarded as minors, 18 to 60 years old are regarded as adults, and those over 60 years old are regarded as seniors, and so, finally, the age of pedestrians is divided into minors, young people, and seniors as the parameters of the scene design. Vehicle speed is normalized to aspect the distance metric calculation of the later algorithm. Continuous variables, namely vehicle speed, were normalized using min–max scaling to transform all features to a common range of [0, 1]. This was performed to prevent variables with larger numerical scales from dominating the Euclidean distance calculation in the subsequent clustering algorithm. The normalization formula is given by

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where x is the original speed value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum speeds observed in the respective dataset.

2.2.2. Hazard Element Selection Based on Random Forest Algorithm

The danger of the scenario is an important factor that cannot be ignored in AEB test scenario studies. The original data recorded pedestrian injuries, which were categorized from high to low as fatal, severely injured, and slightly injured. The Random Forest algorithm is used to quantify the influence of each scene variable on the degree of pedestrian injury to further select the variables used for scene clustering, which can make the scene obtained from clustering more interpretable and distinguishable in terms of danger.

The Random Forest algorithm is a classical integrated supervised learning algorithm based on tree models [12], which performs predictive classification, feature variable selection, etc., by voting on the classification results of multiple decision trees. The main methods for calculating the feature importance of variable-to-label classification based on Random Forest are the information gain method, gain rate, Gini coefficient method and

out-of-bag (OOB) error method. The information gain method and gain rate method are easily affected by the value of variables when judging the importance of variables to labels, and the out-of-bag (OOB) data method calculates the classification error through the k-fold cross-validation method with a long computation time and poorer interpretability than the other methods. The Gini coefficient method is not affected by the value of the variable by calculating the average amount of change in the Gini index of the dataset before and after the classification of the Random Forest variable nodes; therefore, this paper uses the Gini index method to calculate the danger score of the scene elements of the accident data, so as to select the scene elements that have a large impact on pedestrian injuries, which are used for the following clustering scene variables. Using the pedestrian injury level grade as the label of the Random Forest algorithm and the scene elements as the feature variables, the importance score of the scene elements affecting the pedestrian injury level is calculated as the danger score of the scene elements in the following way, so as to quantify the danger of the scene elements.

For any scene element j in the accident dataset after preliminary cleaning, the process of calculating its risk score for pedestrian damage level is shown below:

Step 1: Calculate the Gini coefficient of scene element j before the math node division according to Equation (1), where K is the number of values of variable j , k is the value of variable j , and the probability that variable j takes the value of k is, in this paper, the value of the frequency of taking the value of k .

$$GI_m(D_{clean}) = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (1)$$

Step 2: Calculate the Gini coefficient $GI_m(D_{clean}|j)$ of scene element j after classification at the m th node according to Equation (2); D_1 , D_2 are the two sub-datasets after classification.

$$GI_m(D_{clean}|j) = \frac{D_1}{D_{clean}} GI_m(D_1) + \frac{D_2}{D_{clean}} GI_m(D_2) \quad (2)$$

Step 3: Calculate the average change value VIM_j of the Gini coefficient of the scene elements in the whole Random Forest according to Equation (3), where M is the number of all subcategorized nodes of variable j in a decision tree and n is the number of decision trees in the Random Forest.

$$VIM_j = \frac{1}{n} \sum_{i=1}^n \sum_{m=1}^M (GI_m(D_{clean}|j) - GI_m(D_{clean})) \quad (3)$$

The Random Forest model was implemented using Python3.14's Scikit-learn library (Scikit-learn developers, available at <https://scikit-learn.org>). To ensure robust and interpretable feature importance estimation, the following hyperparameters were configured: the number of trees was set to 500, maximum tree depth was unrestricted to capture complex interactions, and `class_weight = 'balanced'` was used to address the inherent imbalance in injury severity labels. The Gini impurity criterion was chosen over information gain or out-of-bag error due to its computational efficiency and lower sensitivity to variable value ranges, making it suitable for our mixed-type dataset.

Model stability and generalizability were assessed via 5-fold cross-validation. The average out-of-bag accuracy was 0.82 for the straight-road dataset and 0.79 for the intersection dataset, indicating satisfactory predictive performance for the purpose of feature importance ranking. The final importance score for each scene element was computed as the mean decrease in Gini impurity across all trees, normalized to the sum to unity for each dataset. Elements with an importance score below a threshold of 0.05 were considered

to have a negligible impact on injury severity and were excluded from the subsequent clustering phase to reduce dimensionality and enhance cluster interpretability.

In this paper, the above process is realized with the help of Python language and its third-party library, Scikit-learn, to calculate the danger scores of all scene elements and sort them out, and the results are shown in Figures 4 and 5, which show the sorting of the danger scores of the scene elements for the straight-road data and the intersection data, respectively.

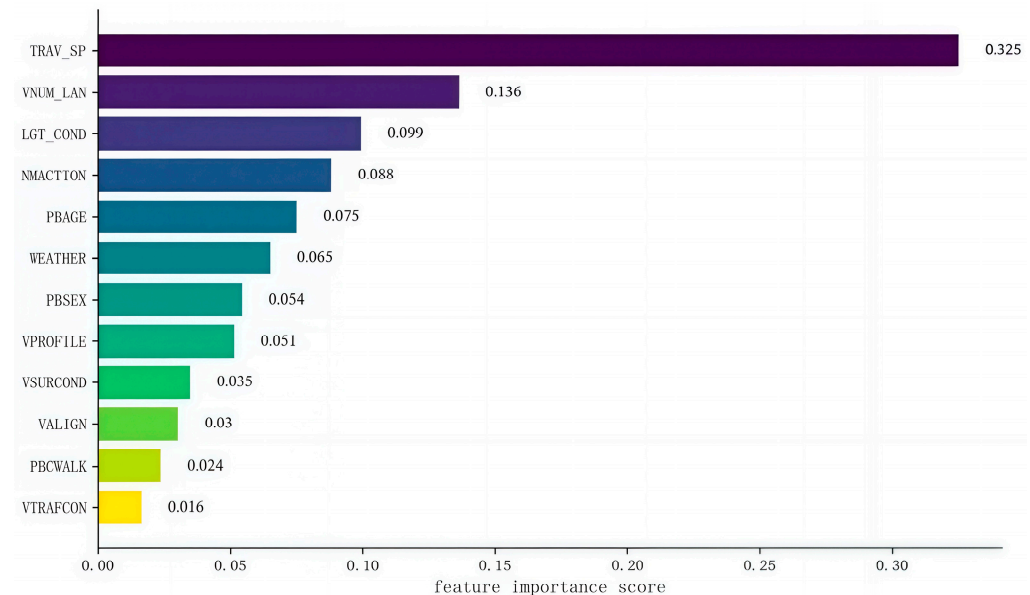


Figure 4. Ranking of importance of elements of the straight-road scenario.

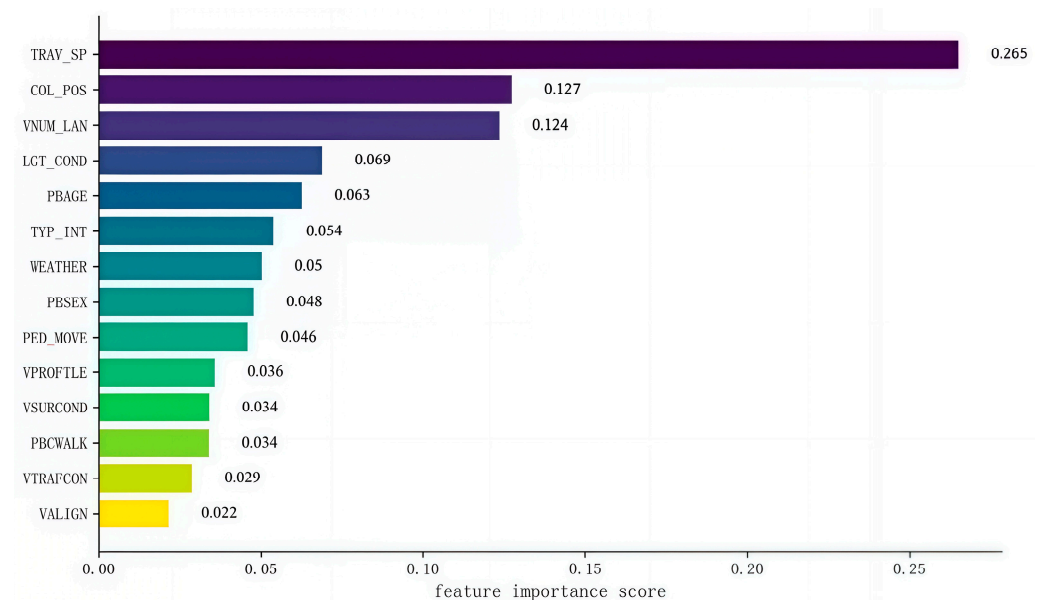


Figure 5. Ranking of importance of junction scene elements.

This section presents the clustering results of pedestrian–vehicle accident data. To enhance interpretability, the importance rankings of scene elements derived from the Random Forest algorithm are visualized in Figures 4 and 5. The *x*-axis of these figures lists the scene elements, while the *y*-axis represents their normalized importance score, where a higher score indicates a greater influence on pedestrian injury severity. All subsequent figures follow a consistent numbering scheme for clarity.

From Figures 4 and 5, it can be seen that the speed of the vehicle has the highest degree of influence on the pedestrian injury level in straight-road and intersection accidents, which is similar to the conclusion obtained from the literature [13–15]. Statistics of the five scenario elements that have the highest degree of influence on pedestrian injuries in straight-road accidents and intersection accidents, in addition to vehicle speed, the four scenario elements that have the highest degree of influence on pedestrian injuries in straight-road accidents are the lane lines of the road, lighting conditions, direction of movement of the pedestrians and the age of the pedestrians, and the four scenario elements that have the highest degree of influence on the pedestrian injuries like the degree of influence on the pedestrian injuries in intersection accidents are the location of the collision, the lane lines of the road, the lighting conditions, and the age of the pedestrians. We can find that four of the five scenario elements that have the highest impact on the severity of pedestrian injuries in intersection accidents versus straight-roadway accidents are the same, indicating that the hazardous scenario elements for straight roadways and intersections are roughly the same. The only difference is the collision location element in intersection accidents, and collision location is the scenario element with the highest impact on pedestrian injury severity other than vehicle speed, indicating that the location complexity of intersection scenarios raises the risk of serious injury or death in intersection pedestrian accidents, which further demonstrates that it is necessary to study the intersection accident data separately from the straight-road accident data.

Based on the importance rankings in Table 1, the four most influential elements for straight-road scenarios—pedestrian movement direction, vehicle speed, lane number, and lighting conditions—were selected as primary clustering variables. Variables such as weather, pedestrian gender, and road surface condition, despite being recorded, were excluded from the clustering process for three reasons. First, their computed importance scores were substantially lower, suggesting a marginal direct effect on injury outcomes in our dataset. Second, preliminary analysis showed low within-scenario variability for these elements. Including low-variance, low-importance features can introduce noise without improving cluster discrimination. Third, these elements are more suitably treated as descriptive parameters rather than discriminative features; they enrich the narrative description of a clustered scenario without defining its core risk logic.

Table 1. Ordering table of elements of the straight-road scenario.

Dimension	Setting Elements	Importance Score
Pedestrian Elements	Direction of pedestrian movement	0.0881
	Pedestrian age	0.075
	Sex of pedestrian	0.0544
Vehicle Elements	Speed	0.325
Transportation	Lane number	0.1363
	Road gradient	0.0514
	Roadmap conditions	0.0348
	Level of road curvature	0.0302
	Availability of crosswalks	0.0236
	Traffic control	0.0163
Environment	Lighting conditions	0.0993
	Weather conditions	0.0651

A sensitivity analysis was conducted by performing clustering with and without these excluded variables. The resulting cluster structures, evaluated using the Adjusted Rand Index, showed a similarity of 0.89, indicating that their exclusion did not fundamentally alter the identified scenario typologies but led to more compact and interpretable clusters.

As can be seen from Table 2, the scene elements that have the highest influence on the pedestrian injury level from the four dimensions of pedestrians, vehicles, traffic environment, and natural environment are the age of pedestrians, vehicle speed, collision location, and lighting conditions, so these four scene elements are selected as the variables for the subsequent clustering, and the type of intersection and the direction of pedestrian movement are also selected as the clustering variables, taking into account the well-established intersection scenarios.

Table 2. Ordering table of junction scene elements.

Dimension	Setting Elements	Importance Score
Pedestrian Elements	Pedestrian age	0.0626
	Sex of pedestrian	0.0477
	Direction of pedestrian movement	0.0459
Vehicle Elements	Speed	0.2652
Transportation	Crash location	0.1273
	Lane number	0.1235
	Intersection type	0.0539
	Road gradient	0.0358
	Roadmap conditions	0.0341
	Availability of crosswalks	0.0339
	Traffic control	0.0287
	Level of road curvature	0.0215
Environment	Lighting conditions	0.0689
	Weather conditions	0.0503

2.2.3. A k-Means Clustering Algorithm Based on Local Outlier Detection

The k-means is one of the most classic and commonly used algorithms in clustering algorithms [16]. It can classify data based on the data distribution characteristics, so it is often used as a means of realizing accident data scenario. The k-means algorithm measures the similarity of the samples in the sample space by calculating the distance between the samples, and commonly used similarity distance measurement formulas are, Euclidean distance, Manhattan distance, Hamming distance, etc. [17]. In this paper, we use the Euclidean distance to measure the similarity of the samples as shown in Equation (4). The features of the sample $x^{(i)}$ and $y^{(i)}$ represent the value of the first feature of any two samples.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x^{(i)} - y^{(i)})^2} \quad (4)$$

The principle of the k-means algorithm is to classify data by continuously optimizing the sum of squares due to error of individual clusters iteratively against random initial clustering centers. The sum of squares due to error (SSE) is obtained by calculating the sum of squares of error from a non-center-of-mass point to the center-of-mass-point in each cluster and summing the sum of squares of error of all clusters cumulatively; the smaller

the value indicates the greater the degree of compactness of clusters, which indicates the better the clustering effect, and its calculation, Formula (5), is shown below:

$$SSE = \sum_{i=1}^k \sum_{j=1}^m \|x_{(i)}^{(j)} - c^{(i)}\|^2 \quad (5)$$

In the equation, k represents that there are k clusters, m represents that there are m samples in the i th cluster, i represents the i th cluster, j represents the j th sample, $x_{(i)}^{(j)}$ represents the j th sample in the i th cluster, and c^i represents the center of the i th cluster.

Road accident data are characterized by high-dimensional nonlinearity and ambiguity. If the k -means clustering algorithm is used directly to cluster road traffic accident data, the clustering results are often easily affected by road accident data noise and the initial clustering center. Therefore, this paper solves this problem by combining a local outlier detection algorithm [18] and hierarchical clustering [19]. Selection of LOF for noise reduction: Among various outlier detection techniques, the Local Outlier Factor (LOF) algorithm was chosen for its ability to identify outliers based on local density deviation. Unlike global methods, LOF is sensitive to local density variations, which is crucial for accident data where “high-risk” or rare event patterns might form sparse clusters within denser, more common accident types. This makes it particularly effective for isolating noise points that are locally inconsistent, thereby improving the quality of the input data for clustering compared to methods like DBSCAN (sensitive to global parameters) or Isolation Forest (better for global anomalies).

Hierarchical clustering does not have to specify the initial clustering center, it determines the similarity between the data points of each category by calculating the distance between them and all the data points, the smaller the distance, the higher the similarity. And the two data points or categories with the closest distance are combined to form a clustering tree. And the distance used to calculate the sample points is also the Euclidean distance.

The local outlier detection algorithm is often used in data mining for outlier detection, which detects whether a sample is an outlier or not by calculating the local outlier score of the sample point to evaluate the density relationship of the sample point in the sample space. For any sample point o in the incident dataset D , the local outlier detection algorithm has the following definition:

k th distance $d_k(o)$: distance to the k th farthest point from the point o .

k th distance field $N_k(o)$: the set of points that are not greater than the k th distance from point o , and the number of points in the set is not less than k .

k th reachable distance $reach_dist(o, p)$: calculated according to Equation (6), where $d(o, p)$ is the distance from point o to point p .

$$reach_dist(o, p) = \max\{d_k(o), d(o, p)\} \quad (6)$$

The parameter k , defining the number of nearest neighbors in the LOF algorithm, was set to 20 after empirical testing across a range of values ($k = 10, 15, 20, 25, 30$). This value represented a balance between locality definition and computational stability. A sample was classified as an outlier if its LOF score exceeded a threshold of 2.0, a common heuristic indicating that the point's local density is approximately half that of its neighbors. This threshold identified approximately 4.2% of straight-road samples and 5.1% of intersection samples as outliers. To assess the sensitivity of our clustering results to these parameters, we repeated the entire pipeline with $k = 15$ and $k = 25$, and thresholds of 1.8 and 2.2. The final cluster assignments showed high consistency, with Jaccard similarity indices above 0.90 across comparisons, confirming that the clustering outcome was robust to minor variations in LOF parameter selection.

The specific procedure for calculating the Local Outlier Factor for any sample point o is as follows:

Step 1: Calculate the distance $d_{(o,p)}$ between point o and the points other than point o in the sample space D according to Equation (4).

Step 2: Determine $d_k(o)$ and $N_k(o)$ of point o by setting the value of k .

Step 3: Calculate the local reachable density $lrd_k(o)$ through Equation (7), where $|N_k(o)|$ denotes the number of points within the k th distance field of point o and p is the point within the k th distance field of point o .

$$lrd_k(o) = \frac{|N_k(o)|}{\sum_{p \in N_k(o)} reach_dist_k(o, p)} \quad (7)$$

Step 4: Calculate the Local Outlier Factor score of the sample point according to Equation (8), which reflects the ratio of the average local reachable density of the points in the k th distance domain of point o to the local reachable density of point o . The closer the value is to 1, indicating that the point o has the same density as the points in its neighborhood, and the smaller the value is compared to 1, indicating that the density of the point o is higher than that of the points in its neighborhood, and the larger the value is compared to 1, indicating that the density of the point o is lower than the density of its neighborhood, the more likely that point o is an outlier.

$$LOF_k(o) = \frac{\sum_{p \in N_k(o)} lrd_k(p)}{|N_k(o)| \times lrd_k(o)} \quad (8)$$

Based on the above theoretical methods outlined, therefore, this paper uses the LOF algorithm to detect outliers in accident data before the clustering analysis of accident data, which performs the hierarchical clustering method to obtain the initial clustering center for the non-outlier dataset, and then uses the k -means clustering algorithm to obtain the clustering results and clustering centers of the non-outlier dataset, and then uses the nearest-distance principle to find the clusters for the non-outlier dataset with the clustering centers. The final clustering result is obtained by clustering division with each clustering center of the non-outlier point set by the principle of nearest distance. In this paper, a k -means clustering algorithm based on the Local Outlier Factor is proposed as follows:

Step 1: Calculate the Local Outlier Factor (LOF) values for each sample in the road accident data based on Euclidean distance and perform sorting.

Step 2: Establish an outlier threshold to filter out the outlier dataset, and apply hierarchical clustering to the non-outlier data for preliminary clustering. Calculate the LOF values for the samples in the preliminary clusters, and select the point with the smallest LOF as the center for the preliminary clustering.

Step 3: Utilize the cluster centers obtained from step two as the initial centers for the k -means clustering. Perform k -means iteration on the non-outlier dataset until convergence. Assign the outlier data to categories based on the nearest proximity principle using the final converged cluster centers and calculate the sum of squared errors (SSE).

Step 4: Calculate the sample means of each cluster as new cluster centers, and repeat step three to assign outliers to categories based on the nearest proximity principle.

Step 5: Compare the SSE; if it decreases, proceed to step five. Continue this process until the SSE reduction reaches a threshold value or the cluster centers no longer change.

2.2.4. Number of Clusters Identification

The selection of the K -value of clustering has a large impact on the effect of clustering, and the general methods for determining the number of clusters for clustering are: the contour coefficient method [20], the elbow rule, the DaviesBouldin index, the AIC, the

BIC and so on [21]. In this paper, the contour coefficient method is used to determine the K-value. The contour coefficient method calculates the contour coefficient by defining the separation and cohesion of the sample, the contour coefficient can measure the clarity of the contour of each category of clustering, and its formula is shown below, which indicates the distance from the sample point to the nearest point in the cluster, called cohesion, and b indicates the distance from the sample point to the nearest point in the nearest other cluster, called separation.

$$S = \frac{a - b}{\max(a, b)} \quad (9)$$

From Equation (9), it can be seen that the contour coefficient of the sample takes a value between -1 and 1 . A smaller value indicates that the sample points are closer to the boundaries of other categories, and when the value is negative it means that they may be assigned to the wrong category. The average contour coefficient is often used to measure the overall clustering effect, where the average contour coefficient is used to select the appropriate K-value.

3. Results

3.1. The Number of Accident Clustering Clusters Is Determined

Considering the size of the dataset, this paper combines the contour coefficient method and the minimum cluster sample capacity to select the K-value. Through the calculation of the dataset, it is found that when the value of K exceeds 10, the minimum cluster sample capacity curve of the straight-road accident data and the intersection accident data will tend to 0, which will lead to too few samples in the clustering cluster difficult to illustrate its typicality, and if discarded it will result in the wasteful use of data. The optimal number of clusters is determined by iterating K from 2 to 10 in order to calculate the average profile coefficient and the sample capacity of the smallest cluster for the clustering results of the straight-road accident dataset and the intersection accident dataset. It can be seen from Figure 6a left that in the straight-road accident data, the average profile coefficient is maximum when $K = 5$ and its minimum cluster sample capacity is greater than 50, so the optimum number of clusters for its clustering is determined to be 5, and the same method can be used to determine the optimum number of clusters for clustering of the intersection accident data to be 7. The effect of the clustering is demonstrated in Figure 6a left and Figure 6b left.

3.2. Analysis of Accident Clustering Results

As described in 2.1 above, K-values of 5 and 7 were selected for clustering analysis of straight-road accident data and intersection accident data respectively. The straight-road accident data are clustered into five types of accident scenarios, of which the third type of accident scenarios have the most samples and the fifth type of accident scenarios have the least samples. Among the 7 types of accident scenarios clustered into the intersection accident data, the 5th type of accident scenario has the most samples and the 6th type of accident scenario has the least samples. In this paper, by the method of taking the value with the largest element frequency as the parameter of the scene elements, we have statistically produced the parameter list of the intersection scene elements in Table 3 and the parameter list of the straight-road scene elements in Table 4.

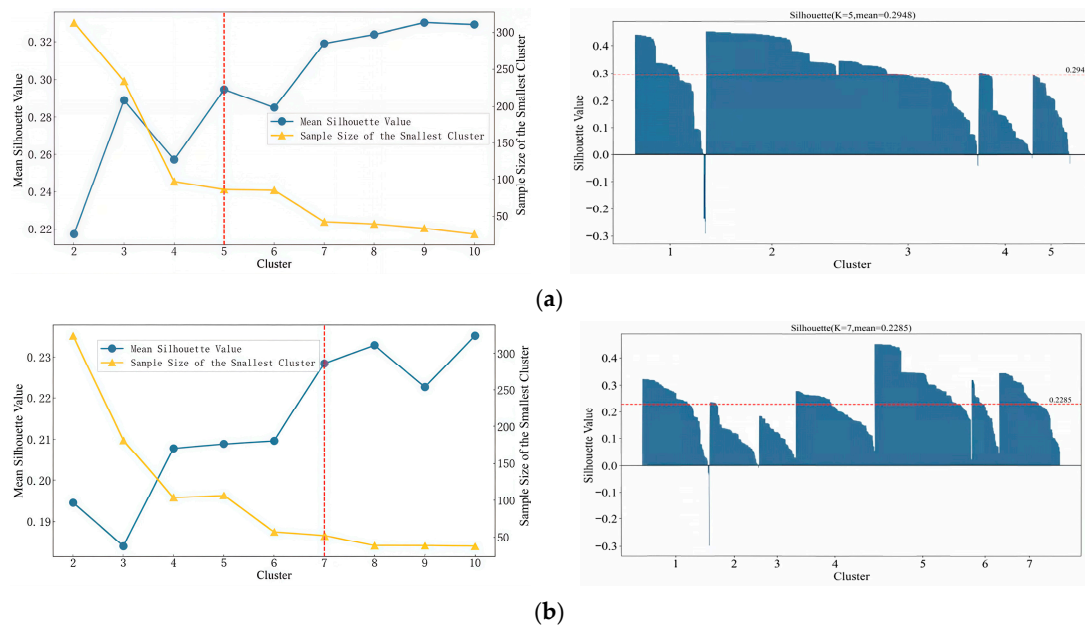


Figure 6. Selection of the number of clusters for clustering based on contour coefficients and minimum cluster sample size (sitting) and sample contour coefficients for $K = 5, 7$ (right). (a) Cluster number selection for straight data clustering. (b) Cluster number selection for intersection data clustering.

Table 3. Dominant parameter values for each clustered straight-road accident scenario.

Features	Form				
	S1	S2	S3	S4	S5
Direction of Pedestrian Movement	Concentric	Downward	Downward	Stationary	Downward
Sex of Pedestrian	Female	Female	Female	Female	Female
Pedestrian Age	Youth	Youth	Youth	Youth	Youth
Lane Number	Two	Two	Two	Two	Six
Road Gradient	Level	Level	Level	Level	Level
Roadmap Conditions	Dry	Dry	Dry	Dry	Dry
Level of Road Curvature	Linear	Linear	Linear	Linear	Linear
Crosswalks	No	No	No	No	No
Traffic Control	No	No	No	No	No
Lighting Conditions	No light at night	Daylight	Light at night	Daylight	No light at night
Weather Conditions	Sunny	Sunny	Sunny	Sunny	Sunny

Table 3 shows that among the characteristic elements, pedestrian gender is female, age is young, road gradient is horizontal, road surface condition is dry, road curvature is straight, no crosswalk, and no traffic signal control have the highest frequency in each type of straight-road scenario. Other elemental parameters in the straight-road scenario of category 1: pedestrian movement in the same direction, the number of lanes is two lanes, and there is no light at night; other elemental parameters in the straight-road scenario of category 2: pedestrians are perpendicular to vehicles, the number of lanes is two lanes, and there is light during the daytime; other elemental parameters in the straight-road scenario of category 3: pedestrians are perpendicular to vehicles, the number of lanes is two lanes, and there is light at night; other elemental parameters in the straight-road scenario of category 4: pedestrians are stationary, the number of lanes is two lanes, and there is light at

night. Elemental parameters: pedestrians are stationary, the number of lanes is two lanes, daytime illumination; and other elemental parameters in the Class 5 straight-road scene: pedestrians are perpendicular to the vehicle, the number of lanes is six lanes, and there is no illumination at night.

Table 4. Dominant parameter values for each clustered intersection accident scenario.

Features	Form						
	J1	J2	J3	J4	J5	J6	J7
Direction of Pedestrian Movement	Driver's left movement	Driver's left movement	Driver's left movement	Driver's right movement	Driver's left movement	Driver's right movement	Driver's left movement
Sex of Pedestrian	Female	Male	Male	Female	Male	Female	Male
Pedestrian Age	Youth	Youth	Youth	Youth	Youth	Youth	Youth
Crash Location	S2	S2	L1	L1	L1	S1	L1
Lane Number	Two	Two	Six	Two	Two	Two	Three/four
Intersection Type	Intersection	T/Y intersection	Roundabout	Intersection	Intersection	T/Y intersection	Intersection
Road Gradient	Level	Level	Level	Level	Level	Level	Level
Roadmap Condition	Dry	Dry	Dry	Dry	Dry	Dry	Dry
Level of Road Curvature	Linear	Linear	Linear	Linear	Linear	Linear	Linear
Crosswalk	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Traffic Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lighting Conditions	No light at night	Good daytime light	Good daytime light	No light at night	Good daytime light	No light at night	Good daytime light
Weather Condition	Sunny	Sunny	Sunny	Sunny	Sunny	Sunny	Sunny

Table 4 shows the parameters of the intersection scene elements. Scenario elements in the pedestrian age for young people, road gradient for horizontal, road surface conditions for dry road surface, road curvature for straight, with pedestrian crossings, with traffic signal control in each type of intersection scenario in the frequency of the largest proportion, so all types of scenarios of the above elemental parameters are the same; in the first category of intersection scenarios in the other elemental parameters: the pedestrian gender is women, the direction of movement is from the driver's left side, the collision location is the S2 area, the number of lanes is two lanes, the intersection type is crossroads, and the lighting condition is nighttime lighting; in the intersection scenarios of category 2, the other element parameters are: the gender of the pedestrian is male, the movement direction is from the driver's left side, the collision location is S2 area, the number of lanes is two lanes, the intersection type is T/Y intersection, and the lighting condition is daytime lighting; in the intersection scenarios of category 3, the other element parameters are: the gender of the pedestrian is male, the movement direction is from the driver's left side, the collision location is for the L1 area, the number of lanes is 6 lanes, the intersection type is roundabout, and the lighting condition is daytime lighting; in the 4th type of intersection scenarios, the other elements of the parameters are: the pedestrian gender is female, the movement direction is from the driver's right side, the collision location is the L1 area, the number of lanes is two lanes, the intersection type is crossroads, the lighting conditions is the nighttime lighting; in class 5 intersection scenarios, the other parameters are: the pedestrian gender is male, the movement direction is from the driver's left side, the collision location is the L1 area, the number of lanes is two lanes, the intersection type is crossroads, and the lighting

conditions are daytime lighting; in class 6 intersection scenarios, the other parameters are: the pedestrian gender is female, the movement direction is from the driver's left side, the collision location is the S1 area, the number of lanes is two lanes, the intersection type is T/Y intersection, and the illumination condition is nighttime illumination; in class 7 intersection scenarios, other element parameters are: the pedestrian gender is male, the movement direction is from the driver's right side, the collision location is the L1 area, the number of lanes is 3/4 lanes, the intersection type is crossroads, and the illumination condition is nighttime illumination.

Figures 7 and 8 show the distribution of crash speeds in the straight-road scenario and the intersection scenario, and the upper and lower quartiles of the speeds in each type of scenario can be found that the speed in the intersection scenario is lower than that of the straight-road scenario, which may be due to the fact that in the intersection scenario, traffic control facilities such as crosswalks, traffic phase signals, and so on, will allow drivers to reduce the speed of the car when passing through, whereas in the straight-road scenario, the road conditions are simple and there are fewer control facilities, so drivers will not pay attention to controlling the reduction in speed. In the straight-road scenario, the road conditions are simple and there are fewer control facilities, so the driver will not pay attention to controlling the speed reduction. Through the analysis of the first chapter of the previous, it can be found that, either in the straight-road or intersection scenes, the collision speed is the most important factor affecting the level of pedestrian collision damage; this paper selects the distribution of the speed of each scene (take the first percentile to the third percentile of the speed interval) of the collision speed parameter. The clustering results show a pronounced prevalence of "female" and "youth" pedestrians across most typical scenarios. This distribution aligns broadly with aggregate U.S. pedestrian crash statistics reported by the NHTSA, which indicate higher exposure or involvement rates for these demographics. However, it is also essential to consider the influence of our data selection criteria. The focus on first-impact, frontal collisions between passenger cars and pedestrians may inherently filter certain accident types. While this selection ensures scenario relevance for current P-AEB system testing, it may underrepresent other demographic groups involved in different crash modalities. Future research could employ stratified sampling or severity-weighted clustering techniques to ensure a more balanced representation of all the vulnerable road user groups in the test scenario library.

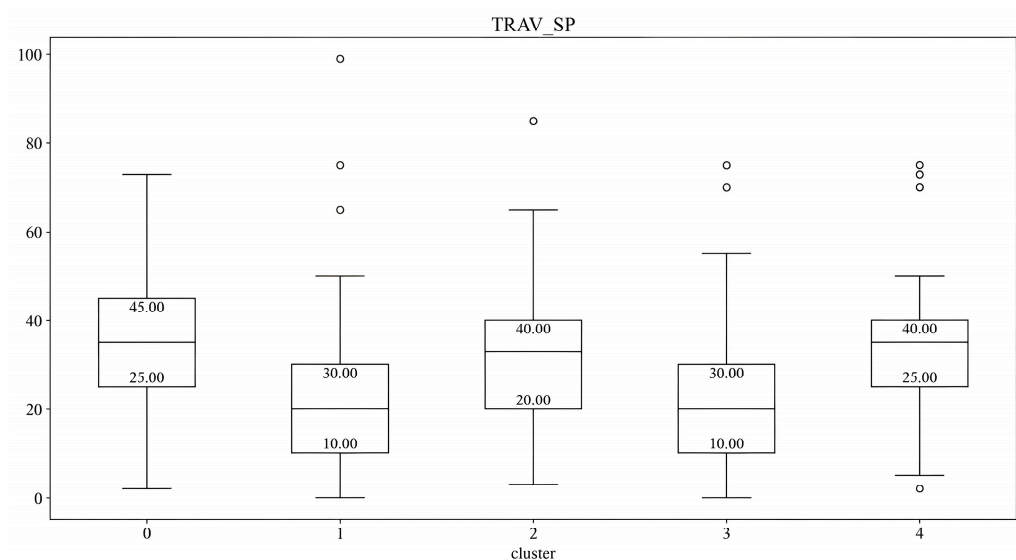


Figure 7. Vehicle speed distribution of various scenarios on a straight-road.

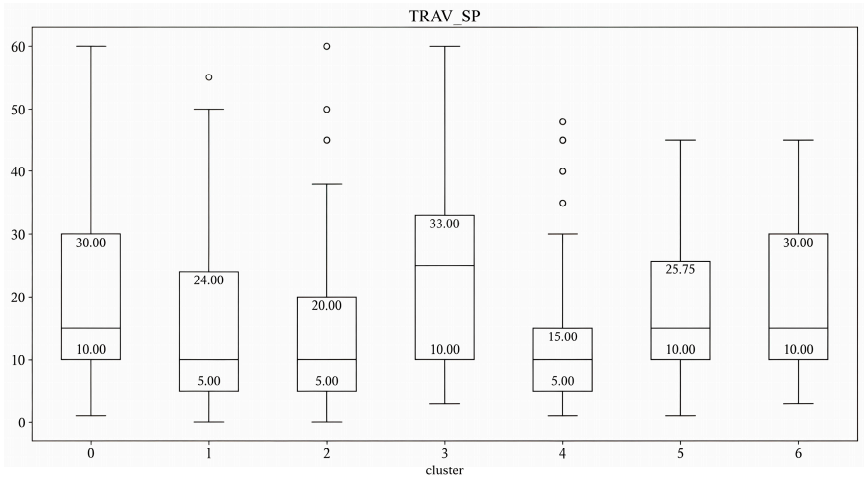


Figure 8. Vehicle speed distribution of various scenarios at junctions.

3.3. Typical P-AEB Test Scenario Design

Based on the five types of typical dangerous collision scenarios obtained from the straight-road scenarios and the seven types of typical dangerous collision scenarios obtained from the intersection scenarios, the statistical collision speed distribution of each scenario can be designed as shown in Table 5 (typical AEB straight-road test scenarios for a human–vehicle system and 7 typical AEB intersection test scenarios for a human-vehicle system). Arrows indicate the direction of movement of vehicles and pedestrians, and circles indicate the collision location.

Table 5. Five types of straight-road test scenarios and seven types of junction test scenarios.

Scene Number	Scene Schematic Diagram	Scene Description
S1		At night without light a passenger car traveling in a two-lane roadway collided with a female youth walking in the same direction
S2		In good daylight a passenger car traveling in a two-lane roadway collided with a female youth walking in a downward direction

Table 5. Cont.

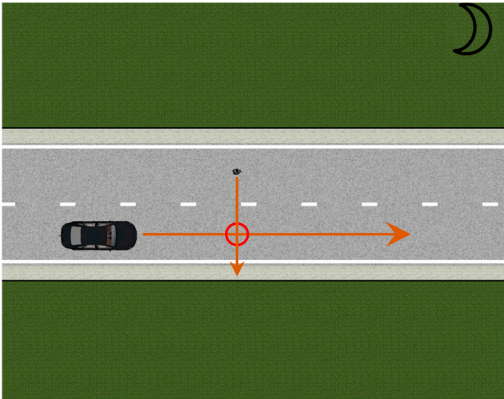
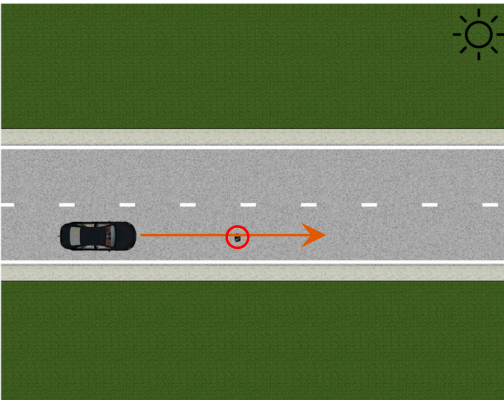
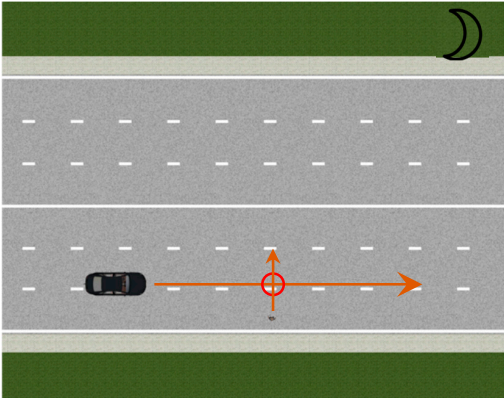
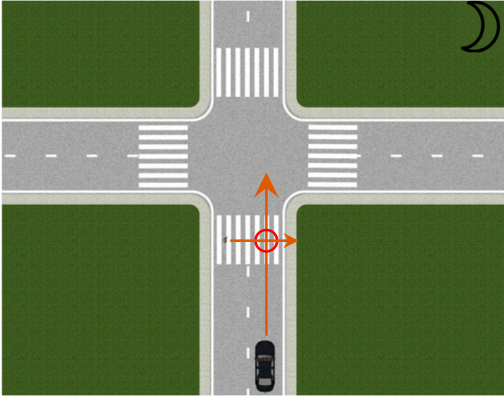
Scene Number	Scene Schematic Diagram	Scene Description
S3		At night with light a passenger car traveling in a two-lane roadway collided with a female youth walking in a perpendicular direction
S4		In good daylight light a passenger car traveling on a two-lane roadway collided with a stationary female youth
S5		At night without light a passenger car traveling in a six lane roadway collided with a female youth walking in a perpendicular direction
J1		At night without light a passenger car traveling on a two-lane roadway at an intersection collides with a female youth coming from the driver's left, collision location S2, speed 10–30 km/h

Table 5. Cont.

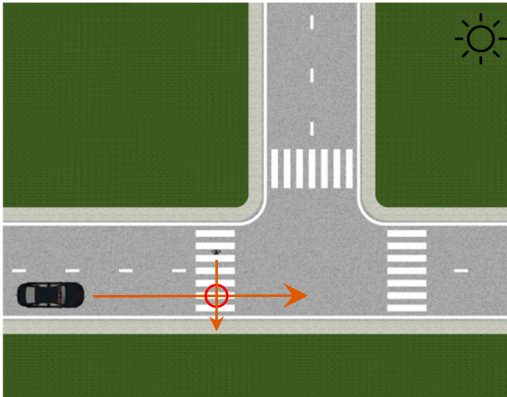
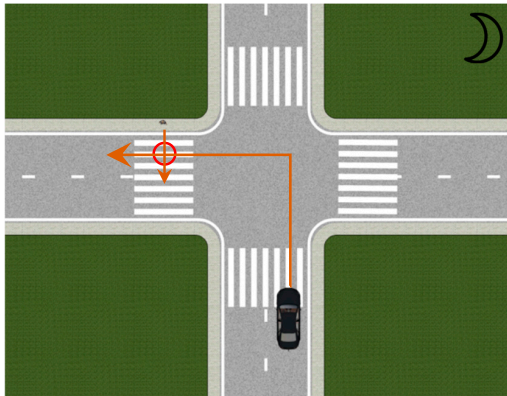
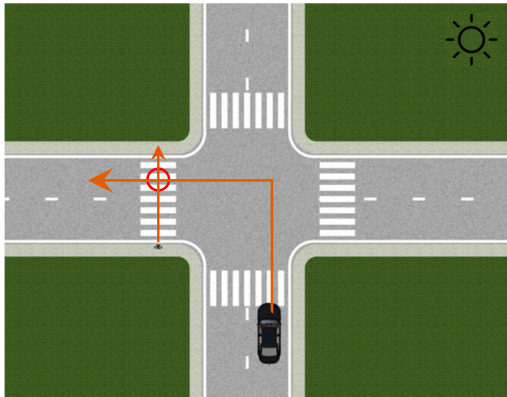
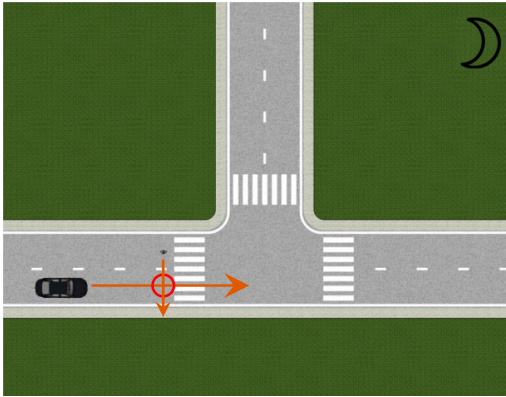
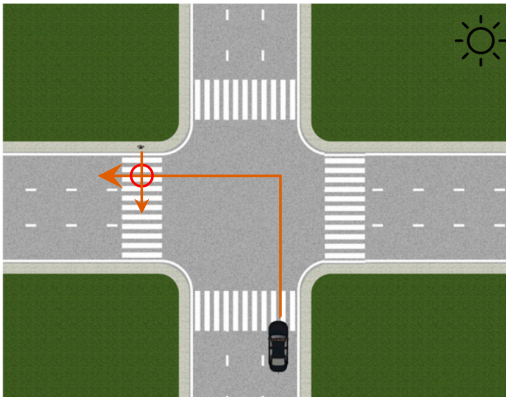
Scene Number	Scene Schematic Diagram	Scene Description
J2		A passenger car traveling on a two-lane T/Y intersection in good daytime light collides with a male youth coming from the driver's left side, collision position S2, speed 5–24 km/h
J3		A passenger car traveling in good light during the daytime collides with a male youth coming from the driver's left side, the collision position is L1 and the speed is 5–20 km/h
J4		A passenger car traveling at night without a license on a cross two-lane road collided with a female youth coming from the right side of the driver, collision position L1, speed 10–33 km/h
J5		In good daytime light a passenger car traveling on a cross two-lane roadway collides with a male youth coming from the driver's left, collision location L1, speed 5–15 km/h

Table 5. Cont.

Scene Number	Scene Schematic Diagram	Scene Description
J6		At night without light a passenger car traveling on a two-lane T/Y intersection collides with a female youth coming from the driver's left, collision position S1, speed 10–26 km/h
J7		In good daytime light a passenger car traveling in three lanes at an intersection collides with a male youth coming from the driver's right, the collision location is L1 and the speed is 10–30 km/h

4. Discussion

The common methods for obtaining test scenarios for intelligent driving functions using road accident data include statistical analysis and clustering analysis, with the latter being able to reduce errors caused by data heterogeneity compared to the former [22]. Feature selection for clustering directly affects the extraction of test scenarios, and selecting features directly related to collisions from accident data as clustering variables is one of the most common approaches [23–25].

4.1. Comparative Analysis with Existing Protocols and Research

To illustrate the practical extension offered by our data-driven approach, the S5 scenario serves as a pertinent example. Current C-NCAP and Euro NCAP protocols for pedestrian crossing primarily address two-lane roads under daylight or illuminated night-time conditions. The S5 scenario, derived from real accident clusters, reveals a significant gap in existing test regimes. It represents a high-risk situation prevalent on urban arterial roads where a pedestrian attempts to cross multiple lanes in low visibility. This scenario critically challenges a P-AEB system's sensor range, late-stage detection algorithms, and decision-making logic, particularly for pedestrians emerging from beyond the immediate lane. Incorporating S5 into the test portfolio would therefore provide a more rigorous assessment of system performance under complex urban and low-visibility operational design domains, enhancing the real-world relevance of safety evaluations.

The extraction of test scenarios from real-world accident data typically follows either statistical or clustering approaches. Compared to purely statistical methods, clustering

techniques such as those used in this study offer advantages in mitigating errors caused by data heterogeneity and uncovering latent structures within complex datasets.

When evaluated against established testing protocols including China C-NCAP and European E-NCAP, the scenarios developed through our data-driven methodology demonstrate both broader coverage and finer granularity. While current regulatory frameworks concentrate on predefined high-frequency scenarios, our approach reveals both common and underrepresented high-risk situations.

For straight-road scenarios, C-NCAP includes standardized tests for adult longitudinal walking and pedestrian transversal crossings. Our results confirm similar scenarios but additionally identify a multi-lane transversal crossing situation absent from existing protocols. This scenario proves particularly relevant for simulating high-speed urban arterials with complex lane configurations.

Regarding intersection scenarios, beyond conventional left-turn and right-turn situations covered by C-NCAP, our clustering identified T/Y-intersection straight-motion conflicts and multi-lane roundabout scenarios. By incorporating collision location as a key clustering variable, our scenarios provide more spatially precise representations of intersection accidents, a dimension frequently oversimplified in standard testing.

Methodologically, this study advances prior clustering-based research by introducing risk-informed feature selection. Earlier studies typically cluster accident data directly using kinematic or geometric variables, while we first employ a Random Forest model to quantify how each scene element influences pedestrian injury severity. This ensures resulting clusters are statistically distinct and intrinsically connected to real-world injury outcomes, thereby enhancing the practical relevance of extracted scenarios for safety system validation.

To quantitatively highlight the extensions offered by our work, Table 6 provides a direct comparison between the scenarios identified in this study and those codified in major international testing protocols.

Table 6. Comparison of extracted scenarios with existing P-AEB test protocols.

Category	Scenario ID (This Study)	Corresponding Scenario in C-NCAP/Euro NCAP	Key Distinctions and Contributions of This Study
Pedestrian	S1 (Longitudinal, night, no light)	Adult longitudinal walking (Day/Night)	Specifies nighttime unlit condition, provides a speed range (20–45 km/h) derived from real data.
Pedestrian	S2, S3 (Transversal, day/night)	Near-/Far-side pedestrian crossing	Confirms standard scenarios but distinguishes lighting conditions (daylight vs. nighttime lit).
Pedestrian	S5 (Transversal, multi-lane, night)	Not covered	Novel scenario: Identifies high-risk crossing on six-lane roads, simulating urban arterials.
Intersection	J1, J4, J5, J7 (Various left-turn conflicts)	Vehicle turning (left/right)	Expands coverage: Includes pedestrians from both driver's left and right, different collision locations (S2, L1), and data-driven speed ranges.

Table 6. Cont.

Category	Scenario ID (This Study)	Corresponding Scenario in C-NCAP/Euro NCAP	Key Distinctions and Contributions of This Study
Intersection	J2, J6 (T/Y-intersection straight)	Not covered	Novel scenario: Covers straight-line collisions at T/Y-type intersections, with/without crosswalk.
Intersection	J3 (Multi-lane roundabout)	Limited/No coverage in protocols	Novel scenario: Represents complex multi-lane roundabout conflicts.

Note: “Not covered” indicates no directly equivalent scenario exists in the latest C-NCAP (2024) or Euro NCAP protocols.

4.2. Parameterization and Completeness of Scenario Description

The typical scenarios summarized in Table 5 represent logical scenarios that define key actors, environmental states, and approximate kinematic ranges without specifying all the parameters needed for immediate simulation. Critical quantitative parameters including initial distance between pedestrian and vehicle, pedestrian movement speed, and precise collision point coordinates remain unspecified in current outputs.

This limitation stems primarily from the nature of the CRSS database, which like many police-reported accident databases contains rich categorical and post-crash information but lacks detailed continuous pre-crash trajectory data. To bridge the gap between logical scenarios and executable test cases, we propose a two-stage parameterization framework.

The first stage involves logical scenario definition, which this study completes by identifying combinations of high-risk factors. The second stage requires concrete parameter assignment, where specific values are given to missing kinematic and spatial parameters. This can be achieved through leveraging complementary datasets such as naturalistic driving studies, defining reasonable parameter ranges to generate multiple concrete scenarios from single logical scenarios, and conducting sensitivity analyses during virtual testing to evaluate system robustness across operational domains.

Thus while presented scenarios are not fully parameterized, they establish a robust data-driven foundation upon which specific executable test scenarios can be efficiently constructed for P-AEB system development and validation.

4.3. Limitations and Regional Adaptability

A primary limitation of this study involves its exclusive reliance on United States accident data. Traffic environments vary significantly across regions due to differences in road infrastructure, traffic regulations, vehicle fleet composition, and cultural driving and pedestrian behaviors. Direct application of extracted scenarios to other regions such as China therefore requires careful consideration.

Key distinctions between United States and Chinese traffic contexts that may influence scenario relevance include road user mix with Chinese urban traffic characterized by higher density and more complex interactions with electric two-wheelers, intersection dynamics where pedestrian crossing behaviors often exhibit lower compliance with traffic signals, road geometry differences in lane widths and intersection designs, and vehicle speed variations between regions.

Despite these differences, the methodological framework proposed remains region-agnostic and transferable. The process of utilizing accident data, selecting risk-critical features via Random Forest, and applying enhanced k-means clustering can be directly applied to region-specific databases. The resulting scenarios would naturally reflect local risk patterns.

For immediate cross-regional application, we recommend a scenario adaptation strategy beginning with priority validation of high-risk scenario types likely universally relevant but requiring local testing, followed by parameter calibration using local traffic data to adjust quantitative parameters especially vehicle speed ranges and initial distances, and concluding with scenario augmentation through the analysis of local accident datasets to supplement region-specific high-risk situations absent from United States data.

Future work will focus on applying this methodology to Chinese accident data to develop complementary region-specific test scenarios, thereby contributing to more globally robust and representative P-AEB evaluation frameworks.

4.4. Study Limitations

While this study provides a data-driven framework for P-AEB test scenario extraction, several limitations should be acknowledged. First, the sole reliance on the U.S. CRSS database means the extracted scenarios inherently reflect U.S.-specific traffic patterns, regulations, and driver/pedestrian behaviors. Direct application to other regions, such as China or Europe, requires careful calibration of parameters like typical speeds and consideration of region-specific elements.

Second, the use of police-reported accident data, while rich in categorical information, lacks the detailed pre-crash kinematic trajectories found in naturalistic driving data. Parameters critical for virtual testing, such as precise pedestrian–vehicle initial distance, time-to-collision, and detailed avoidance maneuvers, are not available and must be supplemented from other sources or defined within reasonable ranges.

Third, the clustering methodology prioritizes frequently occurring accident patterns. Although this ensures the representativeness of the typical scenarios, it may underrepresent rare but severe event types. Future work could employ oversampling techniques or severity-weighted clustering to better capture these high-consequence edge cases.

4.5. Synthesis and Implications

In summary, scenarios extracted in this study encompass those covered by existing protocols while extending beyond them through identification of underrepresented high-risk situations including multi-lane straight-road crossings and complex intersection conflicts. Methodological integration of risk-based feature selection with robust clustering ensures derived scenarios are both representative of real-world accidents and inherently dangerous.

Although current scenarios are presented as logical descriptions lacking full kinematic parameterization, they provide semantically rich and data-grounded foundations for generating executable test cases. By following the proposed two-stage parameterization framework and adapting scenarios to local traffic characteristics, this study's results can effectively guide development, testing, and enhancement of P-AEB systems across varied geographical and operational contexts.

5. Conclusions

This study presents a novel, risk-informed methodology for constructing P-AEB test scenarios grounded in real-world accident data. The primary contributions and findings are summarized as follows.

We developed a robust analytical framework that integrates a Random Forest model for risk-based feature selection with an enhanced k-means clustering algorithm. This framework ensures that the derived scenario typologies are both statistically representative and inherently linked to pedestrian injury severity. The clustering process was fortified by incorporating Local Outlier Factor (LOF) detection and hierarchical clustering initialization

to address data noise and arbitrary centroid selection, common challenges in accident data analysis.

The analysis of U.S. CRSS data revealed vehicle speed as the most critical factor influencing injury outcomes, followed by road geometry (e.g., lane number, collision location) and lighting conditions. Distinctly different risk profiles were confirmed between straight-road and intersection environments, validating their separate treatment.

From the data, we extracted five typical straight-road and seven typical intersection P-AEB test scenarios. These include several high-risk situations, such as multi-lane night-time crossings (S5) and complex roundabout conflicts (J3), which are not prominently featured in existing standardized test protocols like C-NCAP and Euro NCAP, thereby extending the coverage of current evaluation frameworks.

The proposed scenarios offer a more comprehensive and authentic foundation for virtual testing of P-AEB systems. They can guide the development of extended test suites that better reflect the complexity of real-world driving environments, ultimately contributing to the enhancement of active safety system validation and development.

To advance this research, the following steps are recommended: First, applying the same methodology to region-specific accident databases, such as China's CIDAS, to develop locally relevant test scenarios. Second, augmenting these logical scenarios with precise kinematic parameters from naturalistic driving studies to create executable test cases. Finally, integrating vehicle dynamics models, particularly tire-road friction considerations, into the scenario definition to fully capture the cyber-physical interdependencies governing AEB system performance and further improve test realism.

Author Contributions: Conceptualization, Y.L., Q.M. and Z.Y.; Methodology, Y.L., Z.Z., Q.M. and Z.Y.; Validation, Y.L.; Investigation, Z.Z. and Z.Y.; Data curation, Z.Z.; Writing—original draft, Y.L. and Z.Z.; Visualization, Q.M.; Supervision, Z.Z., Q.M. and Z.Y.; Project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Statistical Office. Road Traffic Accident Statistics. 2024. Available online: <https://www.zgtjnj.org/index.aspx> (accessed on 25 January 2025).
2. Najm, W.G.; Smith, J.D. *Development of Crash Imminent Test Scenarios for Integrated Vehicle—Based Safety Systems (IVBSS)*; NHTSA: Washington, DC, USA, 2007. Available online: https://rosap.nhtl.bts.gov/view/dot/8883/dot_8883_DS1.pdf (accessed on 15 August 2025).
3. The European New Car Assessment Program. European New Car Assessment Programme. 2024. Available online: <https://www.euroncap.com/en/for-engineers/protocols/vulnerable-road-user-vru-protection/> (accessed on 25 September 2025).
4. C-NCAP (China New Car Assessment Program) Management Rules (2024 Edition). 2024. Available online: <https://www.c-nicap.org.cn/article-detail/1747900203303780353?type=2> (accessed on 18 April 2025).
5. Lenard, J.; Badea-Romero, A.; Danton, R. Typical pedestrian accident scenarios for the development of autonomous emergency braking test protocols. *Accid. Anal. Prev.* **2014**, *73*, 73–80. [CrossRef] [PubMed]
6. Tan, Z.; Che, Y.; Xiao, L.; Hu, W.; Li, P.; Xu, J. Research of fatal car-to-pedestrian precrash scenarios for the testing of the active safety system in China. *Accid. Anal. Prev.* **2021**, *150*, 105857. [CrossRef] [PubMed]
7. Hu, L.; Lu, T.; Li, G.; Zhang, X.; Cai, H. Automatic generation of intelligent vehicle testing scenarios at intersections based on natural driving datasets. *IEEE Trans. Intell. Veh.* **2023**, *9*, 5448–5460. [CrossRef]
8. Huang, H.; Huang, X.; Zhou, R.; Zhou, H.; Lee, J.J.; Cen, X. Pre-crash scenarios for safety testing of autonomous vehicles: A clustering method for in-depth crash data. *Accid. Anal. Prev.* **2024**, *203*, 107616. [CrossRef] [PubMed]

9. Qian, Y.; Qiu, Y.; Xiao, L.; Hu, W.; Dong, H. Research on test scenarios of aeb pedestrian system based on knowledge and accident data. *Int. J. Veh. Saf.* **2022**, *12*, 322–343. [[CrossRef](#)]
10. Sander, U.; Lubbe, N. The potential of clustering methods to define intersection test scenarios: Assessing real-life performance of aeb. *Accid. Anal. Prev.* **2018**, *113*, 1–11. [[CrossRef](#)] [[PubMed](#)]
11. The National Highway Traffic Safety Administration. Available online: <https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/CRSS/> (accessed on 19 April 2025).
12. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
13. Zhang, G.; Cao, L.; Hu, J.; Yang, K.H. A field data analysis of risk factors affecting the injury risks in vehicle-to-pedestrian crashes. *Ann. Adv. Automot. Med.* **2008**, *52*, 199–214. [[PubMed](#)] [[PubMed Central](#)]
14. Islam, M. An exploratory analysis of the effects of speed limits on pedestrian injury severities in vehicle-pedestrian crashes. *J. Transp. Health* **2023**, *28*, 101561. [[CrossRef](#)]
15. Billah, K.; Sharif, H.O.; Dessouky, S. Analysis of pedestrian–motor vehicle crashes in san antonio, texas. *Sustainability* **2021**, *13*, 6610. [[CrossRef](#)]
16. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [[CrossRef](#)]
17. Anderlucci, L.; Hennig, C. The clustering of categorical data: A comparison of a model-based and a distance-based approach. *Commun. Stat.-Theory Methods* **2014**, *43*, 704–721. [[CrossRef](#)]
18. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. Lof: Identifying density-based local outliers. *ACM SIGMOD Rec.* **2000**, *29*, 93–104. [[CrossRef](#)]
19. Giordani, P.; Ferraro, M.B.; Martella, F. *Quantitative Approaches to Human Behavior*; Springer: Singapore, 2020; Volume 1. [[CrossRef](#)]
20. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
21. Chetouane, N.; Wotawa, F. On the application of clustering for extracting driving scenarios from vehicle data. *Mach. Learn. Appl.* **2022**, *9*, 100377. [[CrossRef](#)]
22. Nitsche, P.; Thomas, P.; Stuetz, R.; Welsh, R. Pre-crash scenarios at road junctions: A clustering method for car crash data. *Accid. Anal. Prev.* **2017**, *107*, 137–151. [[CrossRef](#)] [[PubMed](#)]
23. Pan, D.; Han, Y.; Jin, Q.; Wu, H.; Huang, H. Study of typical electric two-wheelers pre-crash scenarios using k-medoids clustering methodology based on video recordings in China. *Accid. Anal. Prev.* **2021**, *160*, 106320. [[CrossRef](#)] [[PubMed](#)]
24. Sui, B.; Lubbe, N.; Bärgrman, J. A clustering approach to developing car-to-two-wheeler test scenarios for the assessment of automated emergency braking in China using in-depth Chinese crash data. *Accid. Anal. Prev.* **2019**, *132*, 105242. [[CrossRef](#)] [[PubMed](#)]
25. Zhou, R.; Huang, H.; Lee, J.; Huang, X.; Chen, J.; Zhou, H. Identifying typical pre-crash scenarios based on in-depth crash data with deep embedded clustering for autonomous vehicle safety testing. *Accid. Anal. Prev.* **2023**, *191*, 107218. [[CrossRef](#)] [[PubMed](#)]

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.