Received: 30 May 2023 | Revised: 10 September 2023 | Accepted: 25 September 2023

DOI: 10.1002/eng2.12793

RESEARCH ARTICLE

Engineering Reports Open Access WILEY

# An analysis of causative factors for road accidents using partition around medoids and hierarchical clustering techniques

**Pendyala Manasa**[1] | **Pragya Ananth**[1] | **Priyadarshini Natarajan**[2] | **K. Somasundaram**[1] | **E. R. Rajkumar**[2] | **Kattur Soundarapandian Ravichandran**[1] | **Venkatesh Balasubramanian**[2] | **Amir H. Gandomi**[3,4]

[1]Department of Mathematics, Amrita School of Physical Sciences, Amrita Vishwa Vidyapeetham, Coimbatore, India

[2]RBG Labs, Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India

[3]Faculty of Engineering & Information Technology, University of Technology Sydney, Ultimo, New South Wales, Australia

[4]University Research and Innovation Center (EKIK), Obuda University, Budapest, Hungary

**Correspondence**
Amir H. Gandomi, Faculty of Engineering & Information Technology, University of Technology Sydney, Ultimo, NSW, Australia.
Email: gandomi@uts.edu.au

**Abstract**

Insufficient progress in the development of national highways and state highways, coupled with a lack of public awareness regarding road safety, has resulted in prevalent traffic congestion and a high rate of accidents. Understanding the dominant and contributing factors that may influence road traffic accident severity is essential. This study identified the primary causes and the most significant target-specific causative factors for road accident severity. A modified partitioning around medoids model determined the dominant road accident features. These clustering algorithms will extract hidden information from the road accident data and generate new features for our implementation. Then, the proposed method is compared with the other state-of-the-art clustering techniques with three performance metrics: the silhouette coefficient, the Davies–Bouldin index, and the Calinski–Harabasz index. This article's main contribution is analyzing six different scenarios (different angles of the problem) concerning grievous and non-injury accidents. This analysis provides deeper insights into the problem and can assist transport authorities in Tamil Nadu, India, in deriving new rules for road traffic. The output of different scenarios is compared with hierarchical clustering, and the overall clustering of the proposed method is compared with other clustering algorithms. Finally, it is proven that the proposed method outperforms other recently developed techniques.

**KEYWORDS**

accidents, Calinski–Harabasz index, Davies–Bouldin index, DBSCAN, hierarchal clustering, K-means, OPTICS, PAM clustering, road safety, silhouette coefficient

## 1 | INTRODUCTION

Road accidents are a global concern, contributing significantly to mortality, disability, and economic burdens. India, in particular, experiences high traffic fatalities, with at least one in ten global traffic deaths attributed to the country.

The impact of road accidents is far-reaching, affecting victims, families, and the economy due to premature deaths, injuries, disabilities, and lost income potential. Preventing accidents is crucial, but fatalities still occur despite everyone's best efforts. Therefore, data mining techniques, especially clustering algorithms, offer a promising avenue to uncover valuable insights from massive traffic accident datasets. Data mining, also known as knowledge discovery in data, enables the extraction of patterns and essential information from vast datasets. Unlike business intelligence, which focuses on analyzing business data, data mining employs various methods and algorithms to identify relationships and patterns within the data.

In the context of road accidents, data mining can assist in predicting future accident patterns based on historical data. For this research, the authors utilized clustering analysis to make predictions about road accidents, explicitly focusing on fatal accidents. Clustering analysis is one of the techniques used to study the contributory factors of road traffic accidents (RTAs). For identifying the contributing factors of RTAs, different clustering algorithms were proposed in the literature. The K-medoids method[1] is used to determine the critical pre-crash events at T-and four-legged junctions, which can be used to verify the safety of autonomous driving systems. The data set consists of 1056 junction crashes in the UK and resulted in 13 T-junction clusters and 64-legged junction clusters.[1]

The authors studied a cluster analysis of the accident-prone areas in Semarang city to find an area's vulnerability.[2] According to their findings, Semarang's highest level of accidents mainly occurred on weekdays. Data in New Mexico were considered to inspect the injury severity in intersection-related crashes for two-year crashes. The k-means cluster technique was used to cluster the road data. The hierarchical Bayesian random intercept models were developed to identify the contributing factors in every cluster. The findings reveal how the number of crash-level, vehicle/driver-level, and cross-level interactions significantly impact driver injury severity and how these findings help prevent crashes. They examine the understanding of crash potentials among teen drivers using a huge dataset (information on roughly 88,000 respondents) of teen survey data obtained in Texas. Taxicab correspondence analysis was used to analyze the data and discovered that males with provisional or unrestricted licenses are among the highest risk groups.

The authors concentrated on identifying potential factors that may be linked to varying levels of pedestrian injury severity resulting from train-pedestrian collisions (excluding suicides) at highway-rail grade crossings (HRGCs).[3] To conduct their analysis, they utilized 10-year data from the Federal Railroad Administration and employed latent class clustering (LCC) as a method of clustering analysis. Results showed that regardless of the HRGCs' parameters, higher train speed was linked to a higher risk of severe injury. All other factors elevated pedestrian injury severity levels, with differing effects in different clusters.

Traditional statistical models in road safety research have limitations in handling complex datasets, leading researchers to adopt machine learning (ML) approaches. Clustering and classification algorithms like K-means, support vector machines, and decision trees are commonly used for accident severity prediction. However, more comparative analysis and exploration of hierarchical clustering's potential in road safety research must be done. In the related works section, a detailed literature survey has been done. From the literature survey, we found that all the proposed clustering algorithms applied to the entire dataset and then found the grouping based on the homogeneity of the attributes. The main contribution of this article is to analyze six different scenarios (different angles of the problem) in the road accidents dataset, which will help us analyze the more profound insights into the problem and help the transport authorities in Tamil Nadu, India, derive new rules for road traffic. To achieve this, the proposed work analyses causative factors for road accidents in Tamil Nadu using partitioning around medoids (PAM) and hierarchical clustering algorithms, and then it will be compared with other state-of-the-art methods. The flow diagram for the proposed method is given in Figure 1.

The article is organized as follows: In Section 2, we discuss the related works and address the gaps identified. In Section 3, we discuss the methodologies used in this research, such as Gower distance, silhouette width, PAM clustering, and hierarchical clustering. The performance analysis of the proposed methods with others is discussed in Section 5, and finally, a conclusion is given in Section 6.

## 2 | RELATED WORKS

Recently, many researchers have been contributing traditional statistical model-based methods used to predict accident fatality and severity. Some of the conventional statistical model-based techniques are the logit model,[4] logic model approach,[5,6] and ordered probit model[7] to predict accident fatality and severity in terms of independent and dependent accident factors like bad road conditions, weather conditions, lack of traffic indication, drunken and driving,

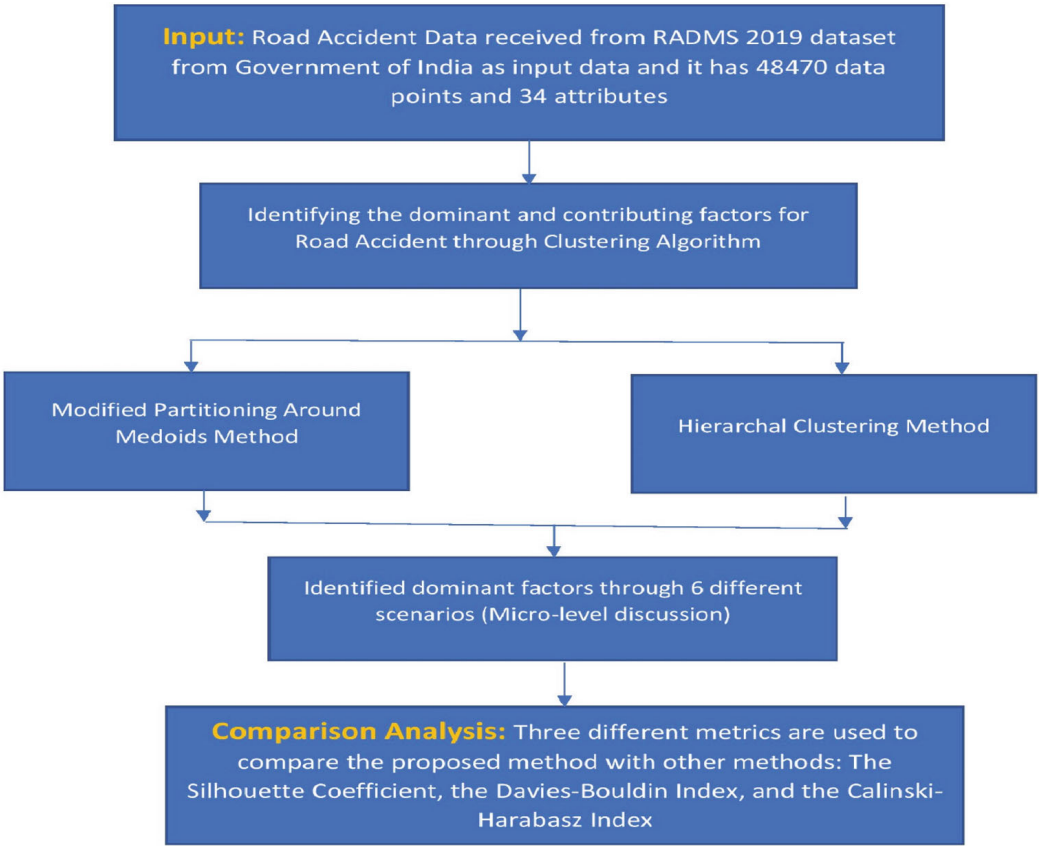



**FIGURE 1** The flow diagram for the proposed method.

children driving, vehicle problem, driver's attitude and so on. The following provides an overview of several studies that utilize clustering and data mining techniques in road safety research. These studies investigate various aspects, such as identifying risky driving behaviors, understanding the relationship between dangerous behaviors and accidents, analyzing injury patterns, and predicting accident causes. Applying clustering and data mining methods in these studies has yielded valuable insights and contributed to developing effective road safety strategies. Lastrucci et al.[8] utilized cluster analysis to identify risky driving behaviors among adolescent drivers in Italy and their association with RTAs. This approach allowed them to identify distinct patterns of risky behaviors and their impact on accident occurrence, providing valuable insights for targeted intervention strategies. Similarly, Hassanzadeh et al.[9] investigated motorcycle riders' riding patterns and risky behaviors in a specific district in Iran using regression analytic methods. This analysis helped them understand the relationships between dangerous behaviors and other factors, contributing to a better understanding of the risk factors involved in motorcycle accidents. Fueyo et al.[10] focused on accident injury patterns, employing unsupervised clustering algorithms on crash data. By classifying seriously injured individuals into clusters, the study opened new possibilities for vehicle safety, potentially leading to improved safety features. In the survey of medical expenses and costs related to motor vehicle crashes in Puerto Rico,[11] K-means clustering played a crucial role in grouping the data, facilitating the identification of the best cluster that maximized distance among groups and minimized distance within groups. This approach contributed to a better understanding of the factors influencing medical expenses and costs associated with injuries in road accidents. Moreover, the survey of data mining methods for road accident analysis[12] presented various clustering and classification methodologies, with the self-organization map (SOM) being used to uncover multiple patterns and predict accident causes. The application of SOM led to improved analysis accuracy compared to k-means clustering, demonstrating the effectiveness of SOM in handling road accident data. In the context of road accidents in Haridwar,[13] India, Sachin Kumar et al. proposed a data mining technique that employed LCC and the k-mode clustering technique to reduce heterogeneity in the dataset. This approach helped reveal crucial facts about the accidents and paved the way for better solutions and targeted interventions. Furthermore, Kim and Yamashita[14] discussed the utility of K-means clustering in safety research and its application in analyzing spatial patterns of pedestrian-involved

crashes in Honolulu. They suggested that both K-means and hierarchical clustering techniques are valuable tools in the arsenal of spatial analytic methods for road safety research. Clustering techniques are not only used in predicting road accident severity, but they can also be used in several other fields like management, arts, engineering, and medicine.[15]

More recently, Sivasankaran and Balasubramanian[26] studied the patterns in road crashes in Tamil Nadu from 2009 to 2017 reported in the Road Accidents Database Management System (RADMS) to explore the injury severity levels of bicycle-vehicle crashes. Latent Class Clustering (LCC) models and binary logit models were combined to identify significant factors in demographics, vehicle, and environmental causes for the crashes. Sivasankaran and Balasubramanian[27] used the same RADMS database to identify associations between pedestrian hit-and-run causes. The same team used Multiple Correspondence Analysis (MCA)[28,29] to identify associations between various contributing factors of pedestrian crashes. Pedestrians of 25-34 age group were associated with crashes at traffic signals where the drivers exhibited non-respect for the right way of rules. In addition, driving violations such as driving against traffic flow and risky driving behaviours such as changing lanes without due care and dangerous overtaking were associated with pedestrian-vehicle crashes.[29] Similar studies were conducted, where the factors associated with the overspeeding risky behaviour of drivers were studied using logistic regression.[30] With a majority of crash fatalities in Tamil Nadu involving motorcycles,[31] ordered logit model was used to identify significant contributing factors in single vehicle motorcycle fatalities.[32] In road safety research, traditional statistical model-based techniques have long been utilized to predict accident fatalities and severity. However, conventional statistical models have limitations, particularly in dealing with complex and multidimensional datasets. Nowadays, most researchers have turned to ML approaches to overcome these challenges due to their predictive superiority, efficiency, and ability to handle informative datasets. The notable works are given as follows: Kwon et al.,[23] used decision trees and Nave Bayes to classify road accidents, and the data were collected between 2004 and 2010; they also compared the classification results with linear regression. Sharma et al.[24] demonstrated the road accident prediction through a support vector machine and multi-layered perceptron; they considered only two parameters, namely, drunken and driving and speed of the vehicle. AlMamlook et al.[25] utilized Nave Bayes, AdaBoost, random forest, and logistic regression methods for road accident predictions. Ester et al.,[26] proposed a density-based algorithm for discovering clusters in large spatial databases with noise (DBSCAN) for clustering, and Ankerst et al.,[27] proposed ordering points to identify the clustering structure (OPTICS) clustering algorithms.

ML has applications in various domains, including construction, occupational accidents, agriculture, education, sentiment analysis, banking, and insurance. Data mining, ML, and deep learning algorithms have been extensively used in road accident prediction. Notable clustering and classification algorithms have been employed to build accident severity models, such as K-means, support vector machines, K-nearest neighbors, decision trees, artificial neural networks, convolutional neural networks, and logistic regression. The literature needs a comprehensive comparative analysis of different clustering algorithms' performance and standardized evaluation metrics. From the literature survey, we found that all the proposed clustering algorithms applied to the entire dataset and then found the clustering based on the homogeneity of the attributes. The significant contribution of this article is to analyze six different scenarios (different angles of the problem) in the road accidents dataset, which will help us analyze the more profound insights into the problem and help the transport authorities in Tamil Nadu, India, derive new rules for road traffic. To achieve this, the proposed work analyses causative factors for road accidents in Tamil Nadu using PAM and hierarchical clustering algorithms, and then it will be compared with other state-of-the-art methods.

## 3 | METHODOLOGY

### 3.1 | Dataset

This article uses the road accident data management system (RADMS) data and GIS-based software for collecting, comparing, and analyzing road accident data for testing. This database is maintained by the State Transport Planning Commission of Tamil Nadu and is the official source that offers complete information on accident circumstances (please refer: https://data.gov.in/catalog/road-accidents-india-2019). Trained police officials compile the crash data across the state with the same instruction manual. The World Health Organization has also advocated using RADMS as an ideal system for nations lacking databases that store accident data. The RADMS 2019 dataset has 48,470 data points and 34 attributes; another dataset contains 2821 data points with 32 attributes.

## 3.2 | Maintaining the integrity of the specifications

Accidents involving grievous injuries and vehicle damage only (non-injury) on national and state highways of Tamil Nadu state in India in 2019 have been considered. Table 1 illustrates the data spread across the various variables for the RADMS dataset.

## 3.3 | Techniques used

Clustering algorithms and techniques play an essential role in analyzing traffic accidents. They can identify groups of people on the road, vehicles, environmental factors, and other such attributes, which would help arrive at conclusions and appropriate countermeasures well.[14]

### 3.3.1 | Gower distance

Choosing the right metric to calculate the distance between two data points, especially while clustering the data, is very important. The RADMS data has both numerical and categorical attributes, hence mixed data. Mixed data have unique metrics for calculating the distance between data points. Gower distance is one such metric used on diverse data.[28] Gower distance is a dissimilarity-based distance metric computed as the mean of partial dissimilarities between data points. In the R programming language used for this study, the daisy function has been used to compute Gower distance. For calculating the Gower distance matrix, the daisy function does the following—each variable (column) or attribute is standardized by subtracting the minimum of the column from each data point and then dividing each data point by the range of the corresponding attribute. This standardization of each variable scales the data such that the range becomes [0, 1]. We compute a measure for each pair of data. If these data are numeric, the measure is the absolute value of the difference divided by the range. If the data is not numeric, the measure takes the value of 1 if the data points are different or 0 if the data points are the same. Gower distance is the average of all these measures.

### 3.3.2 | Silhouette width

Silhouette analysis decides the optimum number of data clusters.[29] The silhouette value describes how similar an object is to its cluster compared to others. The silhouette plot, which represents the same, has the number of clusters on the x-axis and silhouette width on the y-axis, which is given in Figure 2A,B. The higher the silhouette width, the better would be the clustering. silhouette width values lie in the range of −1 and 1. A value of 1 indicates a considerable distance from this sample to its neighboring clusters. A value of 0 indicates that the sample lies on the boundary of two clusters.

The optimum number of clusters is chosen with the help of the silhouette plot and used in the PAM algorithm, which is given in Figure 2 with two data subsets, namely, grievous injury subset (Figure 2A) and no injury subset (Figure 2B). Moreover, a negative value indicates that the sample has been classified into the wrong cluster.

### 3.3.3 | Partition around medoids clustering

The partition around medoids (PAM) clustering algorithm finds objects called medoids around which clusters are built. PAM aims to minimize the average dissimilarity of data points to their closest medoid. The similarity coefficient can evaluate the similarity between the various attributes.[30] If the value of the similarity coefficient is high, then the similarity between the attributes is more elevated. Otherwise, dissimilarity is more significant. In this case, the dissimilarity can be estimated by using the relation $S_{ij} = \left( \frac{a\left[\sin\left(\frac{a\pi}{2n}\right)+d\right]}{\left[a\sin\left(\frac{a\pi}{2n}\right)+b+c+ad\right]} \right)$ for $i \neq j$, where $a$ is the number of attributes which is equally importance between the clusters $i$ and $j$; $b$ is the number of attributes which are required in cluster $i$ and not in $j$; $c$ is the number of attributes which are required in cluster $j$ but not in $i$, $d$ is the number of attributes which is neither needed for cluster $i$ nor $j$, and $n$ is the total number of attributes. Equivalently, the sum of dissimilarities can also be minimized.[31] The algorithm has two phases: a build phase and a swap phase. The "k" medoids are selected during the build phase, and

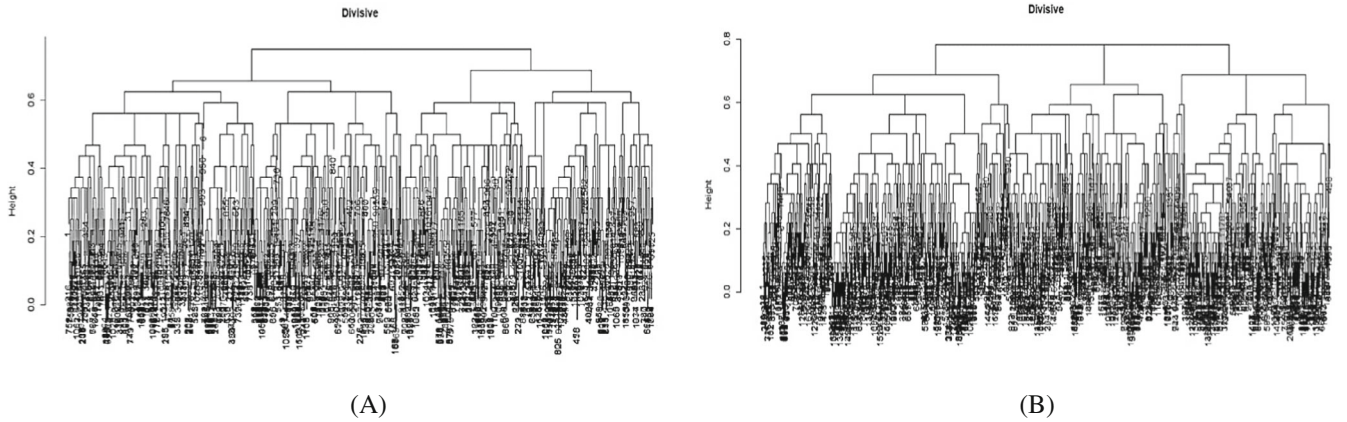
**TABLE 1** Sample data for the road accident data management system (RADMS).

| Category | Frequency | Category | Frequency |
|---|---|---|---|
| Accident severity | | Police present | |
| Grievous injury | 1643 | No | 2108 |
| No injury | 1178 | Yes | 713 |
| Accident day | | Shoulder | |
| Sunday | 456 | Yes | 2821 |
| Monday | 417 | Footpath | |
| Tuesday | 416 | No | 940 |
| Wednesday | 373 | Yes | 1881 |
| Thursday | 366 | Structure narrowing | |
| Friday | 369 | No | 2774 |
| Saturday | 424 | Yes | 47 |
| Collision type | | Other features involved | |
| Head on | 1095 | Tree | 6 |
| Hit pedestrian | 437 | Animals | 17 |
| Hit from rear | 584 | Fixed objects | 29 |
| Hit animal | 11 | Posts | 9 |
| Hit from side | 264 | Advertising boards | 2 |
| Hit tree | 17 | Not applicable | 2758 |
| Sideswipe | 24 | Location type | |
| Skidding | 100 | Corporation | 524 |
| Rain off-road | 16 | Municipality | 968 |
| Overturning | 23 | Panchayat | 1329 |
| Overturning-no collision | 4 | Landmark 1 | |
| Hit object on the road | 38 | Near school/college | 230 |
| Hit object off-road | 25 | Near/inside a village | 255 |
| Hit parked vehicle | 22 | Near factory/industrial area | 85 |
| Others | 161 | Near religious place | 88 |
| Central divider | | Near recreation place/cinema | 23 |
| No | 1045 | In bazaar | 158 |
| Yes | 1776 | Residential area | 114 |
| Junction type | | Near hospital | 116 |
| T-junction | 189 | Open area | 187 |
| Staggered junction | 2 | Near bus stop | 771 |
| Y-junction | 26 | Near petrol pump | 177 |
| Cross junction | 93 | At pedestrian crossing | 23 |
| Roundabout junction | 10 | Narrow bridge or culvert | 17 |
| Junction with more than 4 arms | 29 | Near office complex | 97 |
| Bridge (flyover) | 17 | Near beach | 7 |
| Rail crossing-manned | 4 | Near bridge | 205 |
| Not a junction | 2451 | Junction | 195 |

(Continues)

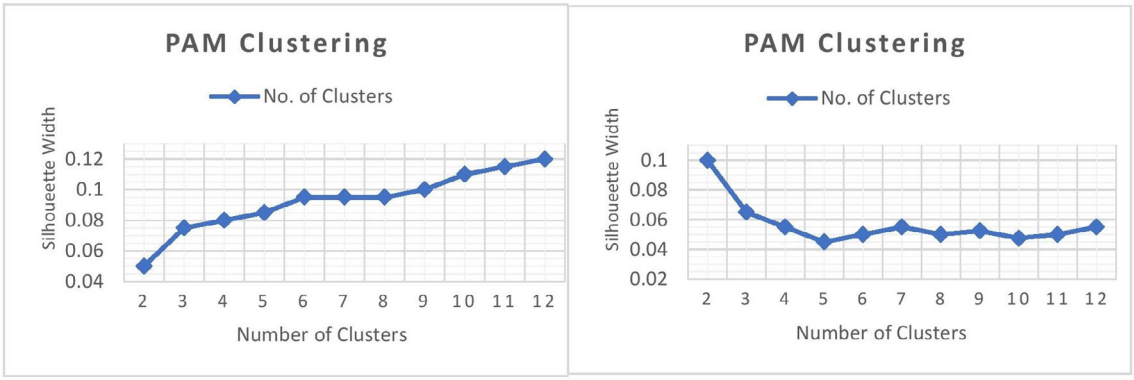
**TABLE 1** (Continued)

| Category | Frequency | Category | Frequency |
|---|---|---|---|
| Junction control | | Near railway station | 10 |
| Not a junction | 2068 | Near traffic signal | 45 |
| Police officer | 16 | Near tollgate | 18 |
| Traffic signals | 46 | Rural–Urban | |
| Flashing signal | 31 | Rural | 2233 |
| Stop sign | 18 | Urban | 588 |
| Give way sign | 42 | Hit run | |
| No control | 600 | No | 2640 |
| Road category | | Yes | 181 |
| National highway | 1054 | Collision type code | |
| State highway | 1767 | Pedestrian accident | 517 |
| Light condition | | Single vehicle accident | 217 |
| Daylight | 1775 | Single vehicle: collision with a fixed object | 29 |
| Twilight | 245 | Vehicles at perpendicular direction without turning | 32 |
| Darkness-no street lights | 174 | Accident between vehicles from opposite direction | 533 |
| Darkness street lights on | 460 | determine factors and identify high-risk | 319 |
| Darkness with poor street light | 167 | Accidents between vehicles from the same direction and where one is turning | 121 |
| Weather code | | Accidents between vehicles from opposite directions and where one is turning | 75 |
| Fine | 2718 | An accident between vehicles in the perpendicular direction and where one is turning | 52 |
| Cloudy | 48 | Accident due to driver error | 926 |
| Light rain | 9 | Collision description code | |
| Heavy rain | 8 | Junction-pedestrian crossing from left to right | 38 |
| Flooding of causeways/rivulets | 2 | Vehicle turning left | 17 |
| Hail/sleet | 1 | Junction-pedestrian crossing from right to left | 25 |
| Smoke/dust | 1 | Pedestrian crossing from left to right | 90 |
| Strong wind | 2 | Pedestrian crossing from right to left | 61 |
| Very cold | 15 | Pedestrian standing on the road | 74 |
| Very hot | 17 | Pedestrian walking along the road | 168 |
| Surface type | | Pedestrian on shoulder/footpath | 61 |
| Tarred (bitumen) | 2300 | Vehicle turning right | 5 |
| Concrete | 372 | Vehicle at a junction | 5 |
| Metaled (WBM) | 37 | Vehicle skidding | 75 |
| Kuttcha | 112 | Loss of control | 100 |
| Road condition | | A passenger fell inside the vehicle | 1 |




(A) (B)

**FIGURE 2** The optimum number of clusters using the silhouette plot for grievous injury and no injury. (A) Grievous injury subset. (B) No injury subset.




(A) (B)

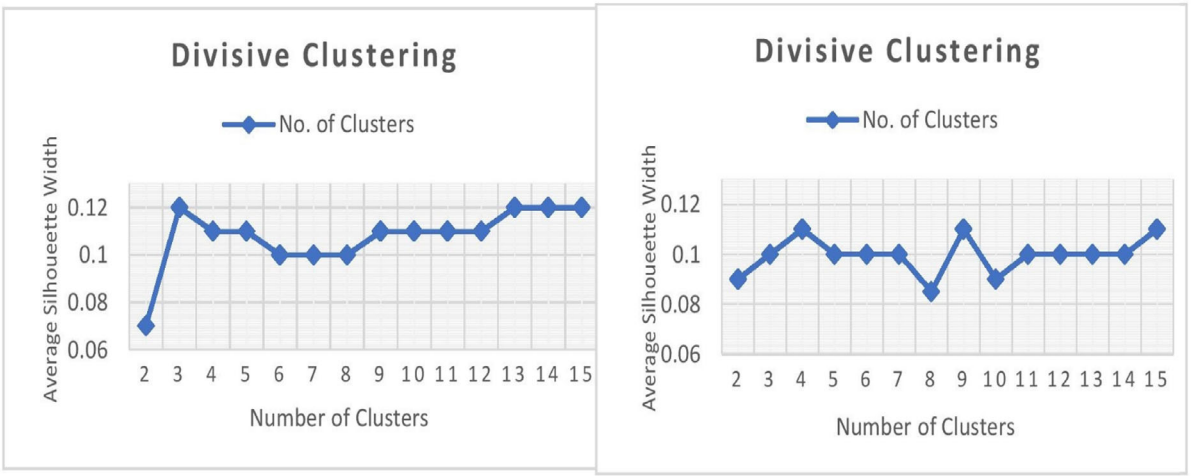
**FIGURE 3** The dendrograms for grievous injury and no injury. (A) Grievous injury subset. (B) No injury subset.

clusters are improved in the swap phase by exchanging selected medoids with better replacements from the non-medoids, if any. It is a more robust version of K-means. PAM clustering is more potent because it accepts a dissimilarity matrix and minimizes the sum of dissimilarities instead of the sum of squares of Euclidean distances. Figure 3 describes dendrograms for the grievous injury (Figure 3A) and the no injury (Figure 3B) cases.

The main advantages of the PAM clustering algorithm over the other clustering algorithms are (i) PAM can effectively deal with noisy and outliers information present in the given dataset, (ii) PAM uses medoid to partition attributes into clusters rather than centroids, and (iii) PAM achieves clustering on overall data rather than on selected samples from the given dataset.

### 3.3.4 | Hierarchical (divisive) clustering

The clustering obtained by using hierarchical clustering consists of two approaches, namely, agglomerative and divisive clustering algorithms. Agglomerative clustering follows a bottom-up approach, where the individual data points are considered as "n" clusters, like a cluster on their own. Then, it finds similarities between them and groups them.[32] All the data points aggregate and form one final cluster in the end. The divisive clustering algorithm follows the top-down approach. The real data is one cluster, divided into sub-clusters until the end of the splits are the data points. Dendrograms are an essential tool that helps decide which of the two approaches in hierarchical clustering can be chosen by gauging the amount of balance/imbalance in the graph. A balanced dendrogram would indicate that that particular algorithm can



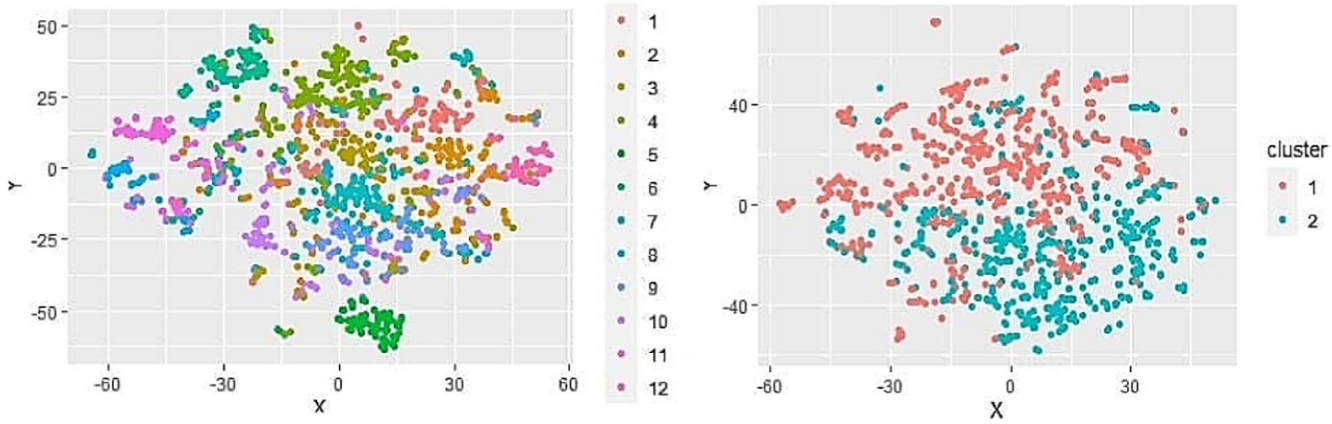

(A) (B)

**FIGURE 4** The hierarchical clustering for grievous injury and no injury. (A) Grievous injury subset. (B) No injury subset.

cluster the data better. Figure 4 explains the number of hierarchical clustering obtained by using the silhouette dataset for the grievous injury (Figure 4A) and the no injury (Figure 4B) cases.

## 4 | PERFORMANCE ANALYSIS OF THE PROPOSED PAM AND HIERARCHICAL CLUSTERING

The results can be split upon analyzing the data into six unique scenarios, four for grievous injuries and two for no injuries subsets. Each scenario consists of PAM clusters and the relevant cluster from Hierarchical clustering, which validates those results. Some scenarios are described in some clusters of PAM, which are unique and not shown by the hierarchical clustering method. All the factors in hierarchical clustering are common to the clusters of both algorithms, but PAM clusters give more details that are not described by the clusters of hierarchical clustering. This makes PAM a more robust algorithm. The total number of accidents with grievous injuries is 1643, and no injuries are 1178. Figure 5 represents the number of PAM clustering obtained by using the silhouette dataset for the grievous injury (Figure 5A) and the no injury (Figure 5B) cases.




(A) (B)

**FIGURE 5** The PAM clusters for grievous injury and no injury. (A) Grievous injury subset. (B) No injury subset.

## 4.1 | Grievous injuries

In this section, we have discussed the results of the unique and utilizing factors in five different scenarios, which are given below:

### 4.1.1 | Scenario 1

"Give way sign present, paved shoulder, flat roads, taking inattentive turn."

PAM clusters in this scenario: Cluster 1 (size = 165 accidents), Cluster 3 (size = 115 accidents), Cluster 10 (size = 101 accidents).

Hierarchical clusters validating this scenario: Cluster 1 (size = 713 accidents) list of uniting factors: Junction control, shoulder type, and road vertical characteristics.

List of unique factors: collision type, road category, traffic restriction, road narrow row, location type, landmark, collision type code, collision description code.

In PAM clustering, two collision types occurred: hit pedestrian (Cluster 1 = 30.9%) and hit from rear (Cluster 3 = 66.08%, Cluster 10 = 48.51%). Accidents involving hitting a pedestrian mostly happen on state highways (75.15%), where the roads are narrow (87.27%), with two-way traffic (81.21%) in the absence of police (95.15%), and heavy vehicles prohibited from entering (67.88%). Specifically, these accidents happened near bus stops (35.15%) in municipality areas (80.61%), and drivers collided with pedestrians who were walking along the road (20.61%). We infer that narrow roads and pedestrians walking along the roads are the misleading use of the accidents here. The presence of police creating awareness among the public to use footpaths, making traffic movement one-way on narrow roads, and initiating road widening activities wherever necessary can be suitable countermeasures to bring down accidents of this type.

Accidents involving collisions from the rear occur on national (Cluster 3 = 80%) and state (Cluster 10 = 78.3%) highways. There was no traffic restriction (58.26%). The national highways were near a panchayat (74.78%) area, and accidents occurred near a bridge (22.61%). Police were not present (82.61%), which leads us to suggest installing some police force, traffic rules, and signages so that entering or leaving the national highway can be smoother and without the risk of any accidents. The state highways were near a municipality (76.24%), and the accidents occurred near a school/college (24.75%). Careless driving (95.05%) was reported to describe the collision. Installing a police force to control and curb careless driving can help reduce these accidents.

In hierarchical clustering, the location was not a junction (81.48%), and if it was, there was not any control present (14.30%), or there was a give way sign (1.40%). The roads had paved shoulders (89.60%) and were flat (97.89%). All these factors are shared between PAM and hierarchical clusters, while PAM clusters further give more details as described above. Table 2 provides an in-depth analysis of Scenario 1: "Give way sign present, paved shoulder, Flat roads, taking inattentive turn" for grievous injury.

### 4.1.2 | Scenario 2

"Traffic signals, darkness with street lights, paved and unpaved shoulders, roads are flat or have a gentle incline, dangerous overtaking, driving against the flow of traffic, and happening in urban areas."

PAM clusters in this scenario: Cluster 6 (size = 125 accidents), Cluster 8 (size = 94 accidents), and Cluster 11 (size = 131 accidents).

Hierarchical clusters validating this scenario: Cluster 2 (size = 463 accidents) list of uniting factors: Junction control, road vertical characteristics, accident cause, rural/urban.

List of unique factors: collision type, light condition, traffic restriction, landmark, collision type code, collision description code.

In PAM clustering, three collision types took place—hit from the side (34.4%), head on (74.5%), and hit from the rear (51.15%). The accidents involving hitting from the side happened in daylight (70.4%), with speed restrictions (88.8%) present, near a traffic signal (17.6%), and the cause reported was careless driving (62.4%). It can be inferred that a sweep from the side could have taken place near the traffic signal despite the restrictions present. A suitable remedy that can be suggested would be to install rumble strips at regular intervals before the signal, as this can help slow vehicles and increase caution.

The accidents involving head-on collisions happened in darkness with street lights on (59.6%), with the entry of heavy vehicles prohibited (78.7%), and in a bazaar area (31.9%). Countermeasures for this situation include widening the

**TABLE 2** In-depth analysis of Scenario 1 (give way sign present, paved shoulder, flat roads, taking in-attentive turn) for grievous injury.

| | HC-1 | PAM-1 | PAM-3 | PAM-10 |
|---|---|---|---|---|
| Uniting factors | | | | |
| Junction control | No control (22.03%) | Not a junction (86.1%) | Not a junction (69.56%) | No control (59.4%) |
| Shoulder type | Paved (89.60%) | Paved (90.3%) | Paved (90.43%) | Unpaved (64.35%) |
| Road vertical characteristics | Flat (97.89%) | Flat (99.39%) | Flat (98.26%) | Flat (97.03%) |
| Unique factors | | | | |
| Collision type | – | Hit pedestrian (30.9%) | Hit from the rear (66.08%) | Hit from rear (48.51%) |
| Road category | – | State highways (75.15%) | National highways (80%) | State highways (78.3%) |
| Traffic restriction | – | Entry of heavy vehicles prohibited (67.88%) | None (58.26%) | None (80.2%) |
| Road narrow | – | Yes (87.27%) | No (86.14%) | No (82.61%) |
| Location type | – | Municipality (80.61%) | Panchayat (74.78%) | Municipality (76.24%) |
| Landmark | – | Near bus stop (35.15%) | Near a bridge (22.61%) | Near a school/college (24.75%) |
| Collision type code | – | Category of pedestrian walking, crossing or standing on/along the road, shoulder or at a junction (42.42%) | Type of vehicle overtaking to the right/left, rear-end collision, when changing the lane to left/right, when making U-turn to the right/left, sight swipe to right/left, sight swipe from opposite direction—right side/left side (53.04%) | Category of rash/careless/drunken driving or disobeying traffic rule (61.4%) |
| Collision description code | – | Pedestrians walking along the road (20.6%) | Rear-end collision (45.22%) | Careless driving (52.48%) |

roads and placing barricades so that speed will automatically be slowed down. The accidents involving hits from the rear happened in daylight (83.2%), with the entry of heavy vehicles prohibited (76.3%) and near bus stops (22.9%). In such a situation, a suitable countermeasure is having designated parking spaces and imposing fines for parking in a no parking area.

In hierarchical clustering, it was found that the road's vertical characteristics were flat (97.40%) or had a gentle incline (1.94%). The accident causes reported were injuries due to human error (67.38%), dangerous overtaking (15.98%), inattentive turn (9.07%), and driving against the flow of traffic (3.45%). Accidents majorly occurred in urban (88.76%) areas where either junction was not involved (71.49%), no control (22.03%) was present at the junction, or there was a traffic signal (3.45%). All these factors are shared between PAM and hierarchical clusters, whereas PAM clusters further give more details as described above. This makes PAM a more robust algorithm. Table 3 provides an in-depth analysis of Scenario 2: "Traffic signals, darkness with street lights, paved and unpaved shoulders, roads are flat or have a gentle incline, dangerous overtaking, driving against the flow of traffic, and happening in urban areas" for grievous injury.

### 4.1.3 | Scenario 3

"Central divider absent, no junction control, paved shoulder, non-respect of rights of way, pedestrians involved, fault of the driver or driver of another vehicle."

PAM clusters in this scenario: Cluster 2 (size = 176 accidents), Cluster 5 (size = 91 accidents), Cluster 9 (size = 125 accidents), Cluster 12 (size = 81 accidents).

**TABLE 3** In-depth analysis of Scenario 2 (traffic signals, darkness with street lights, paved and unpaved shoulders, roads are flat or have a gentle incline, dangerous overtaking, driving against the flow of traffic, and happening in urban areas) for grievous injury.

| Rural/urban | Urban (88.76%) | Urban (99.2%) | Urban (95.74%) | Urban (92.37%) |
|---|---|---|---|---|
| Unique factors | | | | |
| Collision type | – | Hit from side (34.4%) | Head on (74.5%) | Hit from rear (51.15%) |
| Light condition | – | Daylight (70.4%) | Darkness with street lights on (59.6%) | Daylight (83.2%) |
| Traffic restriction | – | Speed restrictions (88.8%) | Entry of heavy vehicles prohibited (78.7%) | Entry of heavy vehicles prohibited (76.3%) |
| Landmark | – | Near a traffic signal (17.6%) | In a bazaar area (31.9%) | Near bus stop (22.9%) |
| Collision type code | – | Category of rash/careless/drunken driving or disobeying traffic rule (74.4%) | Type of collision during overtaking or while making a U-turn/head-collision (59.6%) | Category of vehicle overtaking to the right/left, rear-end collision, when changing the lane to left/right, when making U-turn to the right/left, sight swipe to right/left, sight swipe from opposite direction—right side/left side (48.09%) |
| Collision description code | – | Careless driving (62.4%) | Head on (54.26%) | Rear-end collision (42.75%) |

Hierarchical clusters validating this scenario: Cluster 3 (size = 467 accidents) list of uniting factors: Central divider, junction control, shoulder type, accident cause, contributory factor.

List of unique factors: collision type, road category, location type, traffic movement, traffic restriction, police present, footpath, landmark, collision type code, and collision description code.

Clusters resulting from PAM clustering can be divided into two cases based on the collision type: hit pedestrian (60.23%, 66.67%) in clusters 2 and 12, and head-on collision (96.7%, 72%) in clusters 5 and 9. Accidents involving hitting a pedestrian happened on state highways (67.05%, 60.5%) in panchayat areas (73.3%, 95.1%), with two-way traffic (89.2%, 98.8%) and footpaths present (76.7%, 76.54%). When these accidents took place near bus stops (39.2%), we observed that there was no restriction on traffic (59.66%), and pedestrians were walking along the road (30.68%). Simultaneously, when accidents occurred near a bridge (46.91%), pedestrians were crossing the road from left to right (29.63%). There was a restriction on the entry of heavy vehicles (87.65%). We infer from our findings that there is some inconvenience for pedestrians. Hence, a crossing signal, a traffic signal at the end of the bridge, or a skywalk (before or after the bridge) for the pedestrians to cross can be some countermeasures to mitigate these accidents.

Head-on collisions occurred on national (89%) and state (76%) highways. On a national highway, they came under a municipality area (98.9%). The roads had one-way traffic (98.9%) and did not have footpaths (71.43%). Careless driving (85.71%) was reported as a description of the collision. These accidents happened near a bus stop (34.1%). We infer from this situation that drivers/riders could have been more careful. Considering the state highway accidents, the factors were more alarming. Happening mainly in panchayat areas (76.8%), the roads had two-way traffic (89.6%) with no footpath (98.4%) present and head-on collisions (56%) reported as the collision description. A commonly occurring landmark was near a bus stop (35.2%). Imposing speed limits, fines for violations, widening roads, and building a bus bay wherever necessary can help reduce these accidents. Two-way traffic can be converted into one-way traffic if the roads are very narrow. Installing street lights on narrow main roads of villages can also contribute to reducing these accidents.

Hierarchical clustering found that the roads had a paved shoulder (62.31%) and the central divider was absent (91.01%). Accident sites were not junctions (69.16%); in cases where they were junctions, there was no control (27.19%) at the junction. The accident cause was reported as injured in accidents due to human error (95.29%), and the contributory factor was the fault of the driver/rider (95.93%). All these factors are shared between PAM and hierarchical clusters, whereas PAM clusters further give more details as described above. Table 4 provides an in-depth analysis of Scenario 3: "Central

**TABLE 4** In-depth analysis of Scenario-3 (central divider absent, no junction control, paved shoulder, non-respect of rights of way, pedestrians involved, fault of the driver, or driver of another vehicle) for grievous injury.

| | HC-3 | PAM-2 | PAM-5 | PAM-9 | PAM-12 |
|---|---|---|---|---|---|
| Uniting factors | | | | | |
| Central divider | No (91.01%) | Yes (68.18%) | No (89.01%) | No (96%) | No (87.64%) |
| Junction control | Not a junction (69.16%) | Not a junction (62.5%) | Not a junction (98.9%) | Not a junction (74.4%) | Not a junction (86.42%) |
| Shoulder type | Paved (62.31%) | Paved (78.41%) | Paved (98.9%) | Unpaved (56.8%) | Paved (77.78%) |
| Accident cause | Injured in accidents due to human error (95.29%) | High speed (95.45%) | High speed (86.81%) | High speed (96.8%) | High speed (96.3%) |
| Contributory factor | The fault of the driver/rider (95.93%) | Fault of driver/rider (96.02%) | Aggressive driving (78.02%) | The fault of the driver/rider (96.8%) | Fault of driver/rider (97.53%) |
| Unique factors | | | | | |
| Collision type | – | Hit pedestrian (60.23%) | Head on (96.7%) | Head on (72%) | Hit pedestrian (66.67%) |
| Road category | – | State high ways (67.05%) | National high ways (89%) | State highways (76%) | State highways (60.5%) |
| Location type | – | Panchayat (73.3%) | Municipality (98.9%) | Panchayat (76.8%) | Panchayat (95.1%) |
| Traffic movement | – | Two-way traffic (89.2%) | One-way traffic (98.9%) | Two-way traffic (89.6%) | Two-way traffic (98.8%) |
| Traffic restriction | – | None (59.66%) | None (93.4%) | Entry of heavy vehicles prohibited (62.4%) | Entry of heavy vehicles prohibited (87.65%) |
| Police present | – | Yes (59.1%) | No (95.6%) | No (65.6%) | Yes (80.25%) |
| Footpath | – | Yes (76.7%) | No (71.43%) | No (98.4%) | Yes (76.54%) |
| Landmark | – | Near bus stop (39.2%) | Near bus stop (34.1%) | Near bus stop (35.2%) | Near a bridge (46.91%) |
| Collision type code | – | Category of pedestrian walking, crossing, or standing on/along the road, shoulder, or at a junction (71.02%) | Category of rash/careless/drunken driving or disobeying traffic rule (86.8%) | Category of collision during overtaking or while making a U-turn/head on collision (60.8%) | Category of pedestrian walking, crossing, or standing on/along the road, shoulder, or at a junction (70.37%) |
| Collision description code | – | Pedestrians walking along the road (30.7%) | Careless driving (85.71%) | Head on (56%) | Pedestrians crossing from left to right (29.63%) |

divider absent, no junction control, paved shoulder, non-respect of rights of way, pedestrians involved, fault of the driver or driver of another vehicle" for grievous injury.

#### 4.1.4 | Scenario 4

"Central divider present, head-on collision, near bus stops" PAM clusters in this scenario:

Cluster 4 (size = 223 accidents), Cluster 7 (size = 216 accidents).

List of uniting factors: collision type, central divider, junction control, footpath, landmark.

List of unique factors: traffic restriction, location type, collision description code. This unique scenario is evident only in PAM clustering results and includes 2 clusters with accidents involving head-on collisions (59.64%, 82.87%). These accidents happened when central dividers (95.1%, 78.7%) and footpaths were present (91.03%, 80.56%), police were absent (91.48%, 74.54%), and also mostly near bus stops (30.04%, 38.42%). In municipality areas (61.43%), there was a restriction on entry of heavy vehicles (68.61%), but the collision took place due to rash driving (74.44%). Whereas in panchayat areas (72.68%), there was no traffic restriction (68.05%), and the description of the collision was head-on (68.05%). Based on the whole scenario, the countermeasures can be the presence of police in municipality areas and the imposition of some traffic restrictions in panchayat areas. Table 5 provides an in-depth analysis of Scenario 4: "Central divider present, head-on collision, near bus stops" for grievous injury.

### 4.2 | No injuries

In this section, we have discussed the results with two scenarios under the no injuries category, which are given below:

#### 4.2.1 | Scenario 1

"Central divider present, footpath present, careless driving" PAM clusters in this scenario: Cluster 1 (size = 680 accidents).

Hierarchical clusters validating this scenario: Cluster 1 (size = 671 accidents) list of uniting factors: central divider, traffic movement, footpath, and collision description code.

List of unique factors: collision type, traffic restriction.

The PAM cluster has accidents involving hitting from the rear (30%) and with no traffic restrictions present (47.8%). The hierarchical cluster has accidents involving head-on collisions (30.99%) and hitting from the rear (22.21%), while the traffic restriction prohibited the entry of heavy vehicles (48.28%). The remaining factors gave the same results in both PAM and hierarchical clustering methods, as described here—central divider was present (PAM = 88.23%, HC = 96.42%), the footpath was present (PAM = 84.55%, HC = 94.48%), careless driving reported as collision description code (PAM = 28.1%, HC = 22.35%), two-way traffic (PAM = 81.03%, HC = 82.41%).

**TABLE 5** In-depth analysis of Scenario-4 (central divider present, head-on collision, near bus stops) for grievous injury.

| | PAM-4 | PAM-7 |
|---|---|---|
| Uniting factors | | |
| Collision type | Head on (59.64%) | Head on (82.87%) |
| Central divider | Yes (95.1%) | Yes (78.7%) |
| Junction control | Not a junction po (86.99%) | Not a junction (71.3%) |
| Footpath | Yes (91.03%) | Yes (80.56%) |
| Landmark | Near bus stop (30.04%) | Near bus stop (38.42%) |
| Unique factors | | |
| Traffic restriction | Entry of heavy vehicles prohibited (68.61%) | None (68.05%) |
| Location type | Municipality (61.43%) | Panchayat (72.68%) |
| Collision description code | Rash driving (74.44%) | Head on (68.05%) |

We infer that most of the factors are not leading us toward any causative severe factors, which could imply careless driving. Measures such as driver education during license issues, renewal, or vehicle registration at RTA offices and on hoardings and advertisements are a few countermeasures that could be brought into effect immediately to curb this category of accidents. Table 6 provides an in-depth analysis of Scenario 1: "Central divider present, footpath present, careless driving" for no injury data.

### 4.2.2 | Scenario 2

"Central divider absent, footpath present, head-on collision" PAM clusters in this scenario: Cluster 2 (size = 498 accidents).

Hierarchical clusters validating this scenario: Cluster 2 (size = 507 accidents) list of uniting factors: central divider, traffic restriction, footpath, and collision description code.

List of unique factors: collision type.

The PAM cluster has accidents involving head-on collisions (52%) and with an entry of heavy vehicles prohibited (52.41%).

The hierarchical cluster has accidents involving head-on collisions (42.60%) and hitting from the rear (22.23%), with no traffic restriction (44.97%). The remaining factors give the same results in both PAM and hierarchical clustering methods, as described here—central divider was absent (PAM = 71.5%, HC = 81.26%), footpath was absent (PAM = 72.9%, HC = 85.01%), head-on collision reported as collision description code (PAM = 28.11%, HC = 21.7%), two-way traffic (PAM = 86.75%, HC = 84.81%). We infer that footpaths and dividers are absent, which could be an essential factor responsible for such accident papers. Suitable countermeasures can be to make traffic one-way on roads that are seeing many such accidents and reduce some traffic. Table 7 provides an in-depth analysis of Scenario 2: "Central divider absent, footpath present, head-on collision" for no injury data.

**TABLE 6** In-depth analysis of Scenario-1 (central divider present, footpath present, careless driving) for no injury.

| | HC-1 | PAM-1 |
|---|---|---|
| Uniting factors | | |
| Central divider | Yes (96.42%) | Yes (88.23%) |
| Traffic movement | Two-way traffic (82.41%) | Two-way traffic (81.03%) |
| Footpath | Yes (94.48%) | Yes (84.55%) |
| Collision description code | Careless driving (22.35%) | Careless driving (28.1%) |
| Unique factors | | |
| Collision type | Head on (30.99%), hit from the rear (22.21%) | Hit from the rear (30%) |
| Traffic restriction | Entry of heavy vehicles prohibited (48.28%) | None (47.8%) |

**TABLE 7** In-depth analysis of Scenario-2 (central divider absent, footpath present, head-on collision) for no injury.

| | HC-2 | PAM-2 |
|---|---|---|
| Uniting factors | | |
| Central divider | No (81.26%) | No (71.5%) |
| Traffic movement | Two-way traffic (84.81%) | Two-way traffic (86.75%) |
| Footpath | No (85.01%) | No (71.5%) |
| Collision description code | Head on (21.7%) | Head on (28.11%) |
| Unique factors | | |
| Collision type | Head on (42.60%) hit from rear (22.23%) | Head on (52%) |
| Traffic restriction | None (52.41%) | Entry of heavy vehicles prohibited (52.41%) |

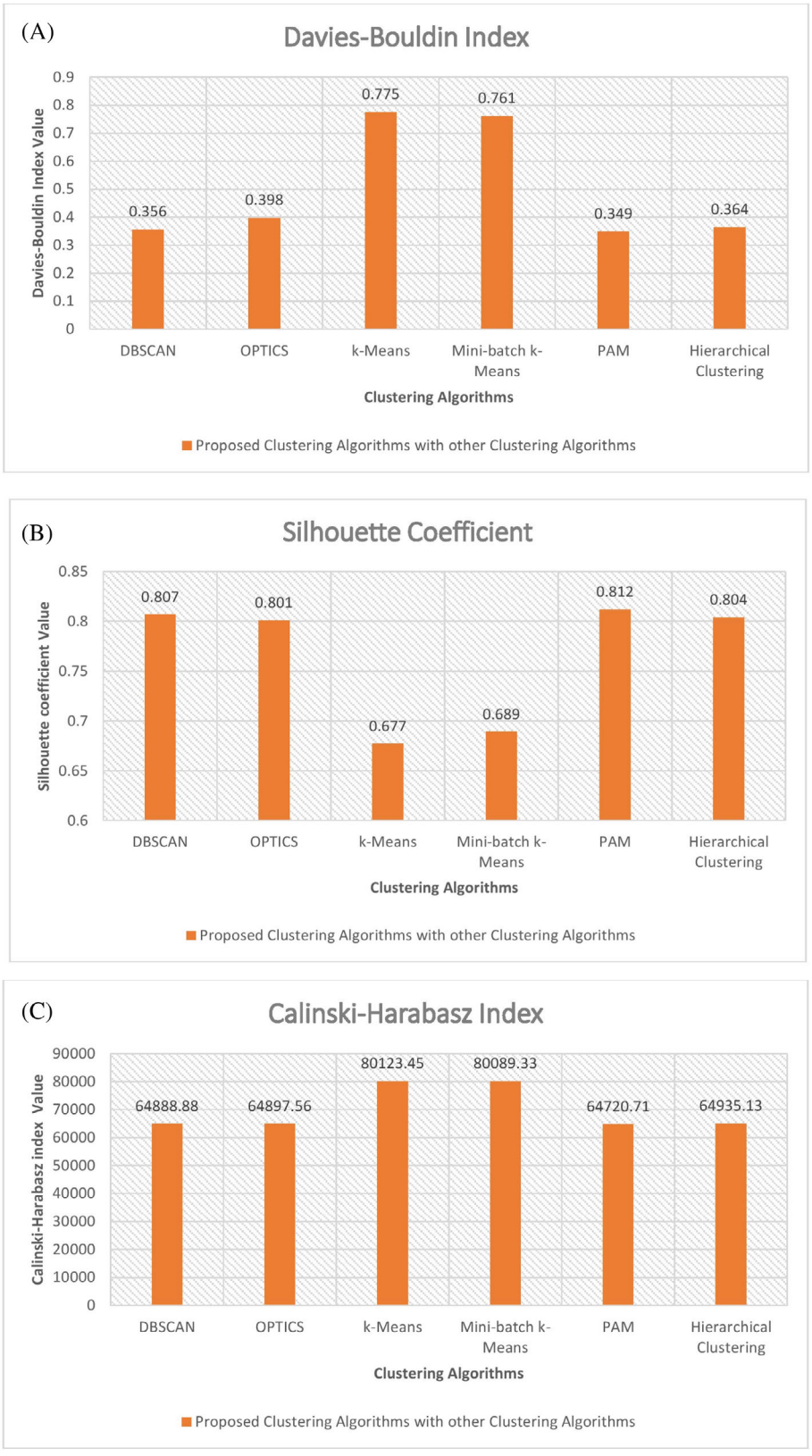



**FIGURE 6** Performance analysis of the proposed method with other methods (A–C). (A) Davies–Bouldin index-based performance analysis of the proposed algorithms with other clustering algorithms. (B) Silhouette coefficient-based performance analysis of the proposed clustering algorithms with other clustering algorithms. (C) Calinski–Harabasz index-based performance analysis of the proposed algorithms with other clustering algorithms.

## 5 | PERFORMANCE ANALYSIS OF THE PROPOSED METHOD WITH OTHER STATE-OF-ART METHODS

The efficiency of the clustering algorithms can be measured by the internal cluster validation metric (ICVM) and the time complexity. In most of the clustering algorithms, the researchers measured only ICVM because it is sufficient to test the performance of the clustering algorithm. There are three performance metrics for evaluating the significance of the clustering algorithms available in the literature: the silhouette coefficient, the Davies–Bouldin index, and the Calinski–Harabasz index. Silhouette coefficient measures usually lie between −1 and +1. It measures how similar an attribute is to attributes in its own cluster compared to attributes in other clusters. Higher, the silhouette value is well matched to its own cluster and poorly matched to other clusters. The Calinski–Harabasz index or variance ratio criterion is the ratio of the sum of inter-cluster and intra-cluster dispersion for all clusters. If the Calinski–Harabasz index is higher, then the performance of the clustering is higher. Davies–Bouldin index is the internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. In contrast to Calinski–Harabasz, the lower the Davis-Bouldin index, the higher the clustering algorithm's performance.

The performance analysis of the proposed method with other methods is given in Figure 6 concerning all three metrics 6A–C.

For the purpose of the performance analysis, PAM and hierarchical clustering (HC) are estimated using the entire dataset. From Figure 6A–C, it is clearly found that the proposed PAM algorithm performs better than the other clustering algorithms.

## 6 | CONCLUSIONS

Tamil Nadu, a state of India, records the highest number of accidents. Compared to other states, Tamil Nadu ranks among the top three in all types of accidents, including those involving fatal, grievous, and mild injuries. Therefore, analyzing Tamil Nadu's accident data and finding countermeasures for every scenario can help Tamil Nadu and other states understand and prevent the current problems that cause accidents beforehand. PAM clustering is a relatively new but robust, hard clustering unsupervised algorithm. It randomly selects medoids from the dataset, calculates distances from data points around them (using a distance measure of our choice), finds a cost, and recalculates these distances as necessary. The algorithm works well with categorical variables, so we choose these from our dataset. Hierarchical clustering is also applied to the same dataset.

Our results show 14 different clusters that fall into six scenarios for our subset of data (accidents with the severity of grievous injuries and vehicle damage only (non-injury) on national and state highways), and we have suggested suitable countermeasures for each scenario. We used the Hierarchical clustering method (divisive approach) to validate the six scenarios' results. Again, the entire dataset has been used to obtain the clustering using PAM and hierarchical clustering, and then these values are compared with other state-of-the-art methods. From the performance analysis, the proposed methodology PAM performs better than the other clustering models. This article uses a novel and robust technique to contribute to solving a national issue of public interest. Our results and countermeasure suggestions will prove beneficial in mitigating rising accidents and saving more lives and property.

One limitation of the proposed PAM clustering algorithm is that it is unsuitable for large datasets due to its high computation requirements. Therefore, our study had to be restricted to a smaller subset of the data. Another limitation is that the algorithm produces new clusters each time it runs. We finalized our clusters after running the algorithm many times and observing a trend in the clusters and silhouette width. We then saved them to a file for further study. This study and algorithm can also be extended and used for any similar purpose involving unsupervised clustering.

### AUTHOR CONTRIBUTIONS

**Pendyala Manasa:** Conceptualization (lead); data curation (equal); methodology (equal). **Pragya Ananth:** Conceptualization (equal); data curation (lead); formal analysis (supporting); writing – original draft (equal). **Priyadarshini Natarajan:** Conceptualization (supporting); formal analysis (supporting); supervision (equal); writing – review and editing (equal). **K. Somasundaram:** Formal analysis (supporting); investigation (lead); methodology (equal); supervision (equal); writing – review and editing (equal). **E. R. Rajkumar:** Methodology (supporting); software (supporting); supervision (equal); validation (equal). **Kattur Soundarapandian Ravichandran:** Formal analysis (supporting); investigation

(lead); supervision (supporting); validation (equal). **Venkatesh Balasubramanian:** Conceptualization (lead); resources (lead); supervision (lead). **Amir H. Gandomi:** Conceptualization (equal); formal analysis (equal); investigation (lead); methodology (supporting); project administration (supporting); supervision (equal); validation (lead); writing – original draft (lead); writing – review and editing (equal).

## ACKNOWLEDGMENT

The authors wish to thank Tamil Nadu Police for providing permission and access to use the Road Accident Data Management System (RADMS) in this study. The authors gratefully acknowledge the Amrita ICTS Department, Amrita Vishwa Vidyapeetham, Coimbatore, for providing a high-performance computing facility for testing our algorithms. Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

## CONFLICT OF INTEREST STATEMENT

No conflict of interest exists in our article.

## PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/eng2.12793.

## DATA AVAILABILITY STATEMENT

https://morth.nic.in/road-accident-in-india

## ETHICS STATEMENT

No ethical issues related to our work because our research involves anonymized records.

## ORCID

*Priyadarshini Natarajan* https://orcid.org/0000-0002-8757-1978
*Kattur Soundarapandian Ravichandran* https://orcid.org/0000-0003-2397-461X
*Amir H. Gandomi* https://orcid.org/0000-0002-2798-0104

## REFERENCES

1. 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.
2. Budiawan W, Purwanggono B. Clustering analysis of traffic accident in Semarang City. *E3S Web of Conferences*. Vol 73. EDP Sciences; 2018:12001.
3. Zhao S, Iranitalab A, Khattak AJ. A clustering approach to injury severity in pedestrian-train crashes at highway-rail grade crossings. *J Transp Saf Secur*. 2019;11(3):305-322.
4. Bedard M, Guyatt GH, Stones MJ, Hirdes JP. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accid Anal Prev*. 2002;34(6):717-727.
5. Haleem K, Alluri P, Gan A. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. *Accid Anal Prev*. 2015;81:14-23.
6. Kim JK, Ulfarsson GF, Shankar VN, Mannering FL. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accid Anal Prev*. 2010;42(6):1751-1758.
7. Zajac SS, Ivan JN. Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural connecticut. *Accid Anal Prev*. 2003;35(3):369-379.
8. Lastrucci V, Innocenti F, Lorini C, et al. Patterns of risky driving behaviors among Tuscan adolescent drivers: a cluster analysis. *Eur J Public Health*. 2020;30(Supplement-5):ckaa165-1111.
9. Hassanzadeh K, Salarilak S, Sadeghi-Bazargani H, Golestani M. Motorcyclist risky riding behaviors and its predictors in an Iranian population. *J Inj Violence Res*. 2020;12(2):161-170.
10. Fueyo S-d, Rocio MJ, Francisco Lopez-Valdes H, Gabler C, Woerner L, Hiermaier S. Cluster analysis of seriously injured occupants in motor vehicle crashes. *Accid Anal Prev*. 2021;151:105787.
11. Bianchi Santiago JD, Didier V, Macchiavelli R. KABCO severity cost estimation by cluster analysis for injury-only crashes in Puerto Rico. No. TRBAM-21-01782; 2021.
12. Sakhare AV, Kasbe PS. A review on road accident data analysis using data mining techniques. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE; 2017:1-5.

13. Kumar S, Toshniwal D, Parida M. A comparative analysis of heterogeneity in road accident data using data mining techniques. *Evol Syst*. 2017;8(2):147-155.
14. Kim K, Yamashita EY. Using a k-means clustering algorithm to examine patterns of pedestrian-involved crashes in Honolulu, Hawaii. *J Adv Transp*. 2007;41(1):69-89.
15. Chandramohan D, Dumka A, Jayakumar L. 2M2C-R2ED: multi-metric cooperative clustering based routing for energy efficient data dissemination in green-VANETs. *Technol Econ Smart Grids Sustain Energy*. 2020;5:15. doi:10.1007/s40866-020-00086-4
16. Sivasankaran SK, Balasubramanian V. Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in India. *Journal of Safety Research*. 2020a;72:127-138. doi:10.1016/j.jsr.2019.12.012
17. Sivasankaran SK, Balasubramanian V. Investigation of factors contributing to pedestrian hit-and-run crashes in India. *Journal of Transportation Safety & Security*. 2020b;14(3):382-403. doi:10.1080/19439962.2020.1781313
18. Sivasankaran SK, Balasubramanian V. Investigation of pedestrian crashes using multiple correspondence analysis in India. *International Journal of Injury Control and Safety Promotion*. 2019;27(2):144-155. doi:10.1080/17457300.2019.1681005
19. Natarajan P, Sivasankaran SK, Balasubramanian V. Identification of contributing factors in vehicle pedestrian crashes in chennai using multiple correspondence analysis. *Transportation Research Procedia*. 2020;48:3486-3495. doi:10.1016/j.trpro.2020.08.104
20. Balasubramanian V, Sivasankaran SK. Analysis of factors associated with exceeding lawful speed traffic violations in Indian metropolitan city. *Journal of Transportation Safety & Security*. 2019;13(2):206-222. doi:10.1080/19439962.2019.1626962
21. Sivasankaran SK, Rangam HK, Balasubramanian V. Injury profiles and epidemiology of single vehicle motorcycle fatalities in Tamil Nadu, India, 2009-2017. *Journal of Road Safety*. 2022;33(3):40-54. doi:10.33492/jrs-d-20-00125
22. Sivasankaran SK, Rangam H, Balasubramanian V. Investigation of factors contributing to injury severity in single vehicle motorcycle crashes in India. *International Journal of Injury Control and Safety Promotion*. 2021;28(2):243-254. doi:10.1080/17457300.2021.1908367
23. Kwon OH, Rhee W, Yoon Y. Application of classification algorithms for analysis of road safety risk factor dependencies. *Accid Anal Prev*. 2015;75:1-15.
24. Sharma B, Katiyar VK, Kumar K. Traffic accident prediction model using support vector machines with gaussian kernel. In: Pant M, Deep K, Bansal J, Nagar A, Das K, eds. *Proceedings of the Fifth International Conference on Soft Computing for Problem-Solving*. Springer; 2016:1-10.
25. AlMamlook RE, Kwayu KM, Alkasisbeh MR, Frefer AA. Comparison of machine learning algorithms for predicting traffic accident severity. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE; 2019:272-276.
26. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. ACM; 1996:226-231.
27. Ankerst M, Breunig MM, Kriegel H-P, Sander J. OPTICS: ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. ACM; 1999:49-60.
28. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971;27:857-871.
29. Batool F, Hennig C. Clustering with the average silhouette width. *Comput Stat Data Anal*. 2021;158:107190.
30. Ravichandran KS, Rao KCS, Saravanan R. The role of fuzzy and genetic algorithms in part family formation and sequence optimisation for flexible manufacturing systems. *Int J Adv Manuf Technol*. 2002;19:879-888. doi:10.1007/s001700200100
31. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley Sons; 2009.
32. Xiong T, Wang S, Mayers A, Monga E. DHCC: divisive hierarchical clustering of categorical data. *Data Min Knowl Discov*. 2012;24(1):103-135.

**How to cite this article:** Manasa P, Ananth P, Natarajan P, et al. An analysis of causative factors for road accidents using partition around medoids and hierarchical clustering techniques. *Engineering Reports*. 2024;6(6):e12793. doi: 10.1002/eng2.12793