ICIC Express Letters
Part B: Applications
Volume **16**, Number **4**, April **2025**

ICIC International ©2025 ISSN 2185-2766
pp. **359–367**

IDENTIFICATION OF TRAFFIC ACCIDENT PRONE AREAS USING DATA MINING

Amil Ahmad Ilham*, Anugrayani Bustamin and Fitriani Nasir

Department of Informatics
Universitas Hasanuddin
Jalan Poros Malino Km. 6, Gowa 92171, Indonesia
anugrayani@unhas.ac.id; nasirfi17d@student.unhas.ac.id
*Correponding author: amil@unhas.ac.id

Received July 2024; accepted September 2024

Abstract. *A traffic accident is a road incident involving vehicles which might result in human casualties or property losses. According to World Health Organization, traffic accident is the leading cause of death for children and adults aged 5-25 years old. In this study, we identified and analyzed accident-prone areas in Makassar City, Indonesia, based on traffic accident case similarity using dataset which consists of 3558 daily traffic accident history data. We developed a case clustering system using unsupervised Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm, whose hyperparameter was determined using a heuristic approach, namely ordered k-dist graph with Nearest Neighbors Searching. The system produces 16 optimal clusters as indicated by a fairly small Sum Square Error value of 71.22 and a Silhouette Score value that is close to 1, namely 0.805. This shows that the resulting clusters have quite close intra-cluster cohesion and have quite far distances between clusters.*
Keywords: Clustering, DBSCAN, K-Means, Traffic accident, WebGIS

1. **Introduction.** Road traffic injuries are now the leading killer of people aged 5-29 years [1]. This condition became a big challenge to the sustainable development targets as stated in Sustainable Development Goals (SDGs) 2030 Target 3.6, which states that the target of reducing the global traffic accident mortality rate is 50% by 2020 [2]. The accidents could result in death victims, serious injuries, minor injuries and material losses. Indirect cost, such as loss of productivity, damage to vehicles and property, reduces quality of life and other affected factors [3]. Based on WHO statistics, developing countries are far more affected by traffic accidents than developed countries, where 75% of deaths due to traffic accidents occur in developing countries, even though these countries only have 32% of the world's motorized vehicles [4].

Several studies have provided a knowledge representation for prone accident issues by using machine learning. DBSCAN and Ordering Points to Identify the Clustering Structure (OPTICS) algorithms were recommended for clustering accident-prone areas to improve effectiveness and efficiency [5]. The K-Means algorithm was used to classify the level of accident vulnerability on Jakarta Bogor Highway to identify certain accident-prone points [6]. Fuzzy C-Means algorithm was used to cluster traffic accidents in the city of Medan, Indonesia to identify vulnerable roads [7].

In our research, we applied data mining clustering techniques to dividing the set of accident cases into several clusters based on their similarities. Data mining, often known as Data Discovery or Knowledge Discovery from Databases (KDD), involves the analysis of data from different perspectives to transform it into valuable information [8]. Clustering, within the realm of data mining, is a method that categorizes items into multiple groups

DOI: 10.24507/icicelb.16.04.359

determined by their proximity or the distance between them [9]. The requirements for ideal clustering are that the distance between clusters must be minimized and the distance between clusters must be maximized [10]. DBSCAN is one of the clustering algorithms, which relies on the concept of density-based clusters and leads to the identification of arbitrary cluster shapes. DBSCAN can be considered a more effective method in finding clusters with arbitrary shapes [11]. Zhang et al. in 2023 proposed a modification of the traditional DBSCAN algorithm to identify road accidents in China [12]. Gershtein and Terekhov applied DBSCAN to identifying severe traffic accident hotspots in Newton, MA and Massachusetts, USA, and found that the algorithm was effective in this context [13].

Dimensionality reduction and geocoding are also required. Principal Component Analysis (PCA) is a dimension reduction technique used to maintain variation between many variables by restating these variables into several new variables, so that most of the variance is still accounted for or maintained by several new, uncorrelated variables [14]. Meanwhile, to provide geographic coordinates that correspond to the location of each cluster, geocoding is carried out using the geopy service and visualized in WebGIS. The study using GIS also provided innovative technology for spatial data modeling to categorize and predict food self-sufficiency zones [15]. Geocoding is the process of digitally translating street addresses into longitude and latitude coordinates on the Earth's surface [16]. To increase the predictive power of the model, predictive geospatial analysis used Principal Component Regression (PCR) which was a combination of PCA and Multiple Linear Regression (MLR) approaches [17].

Our research contributes to the identification of accident-prone areas by utilizing visualization of clustering results in the form of digital maps. The search for optimal parameters was analyzed using a heuristic approach, especially in determining the epsilon value and minimum point in the DBSCAN algorithm and the results were compared to the clustering results using the K-Means algorithm. The clusters produced by the DBSCAN algorithm have quite close intra-cluster cohesion and have quite far distances between clusters compared to the K-Means algorithm. This work is divided into four sections: Section 2 describes materials and method, Section 3 presents results and discussion, and Section 4 contains our conclusions.

2. **Materials and Method.**

2.1. **Dataset.** The dataset used in this research is historical traffic accident data in Makassar City, Indonesia, totaling 3558 historical traffic accident data. The features contained in the dataset are day, time, address, area, type of accident, victim vehicle and suspect vehicle. Accident type and vehicle type codes are presented in Table 1 and Table 2, respectively.

2.2. **Methodology.** We have designed a system consisting of cluster formation and WebGIS visualization as shown in Figure 1.

Table 1. Accident type codes

No	Code	Accident type
1	SVA	single-vehicle
2	HCA	head-on collision
3	BSA	bump into someone
4	HRA	hit and run
5	OCA	out of control
6	CRA	chain reaction
7	MsA	miscellaneous

TABLE 2. Vehicle type codes

No	Code	Vehicle type
1	McV	motorcycle
2	MRV	motor rickshaw
3	PCV	passenger car
4	TRV	trucks with 6 or more wheels
5	CTV	microbus/city transportation
6	BsV	bus
7	PbV	pedicab
8	PRV	passenger motor rickshaw
9	BcV	bicycle

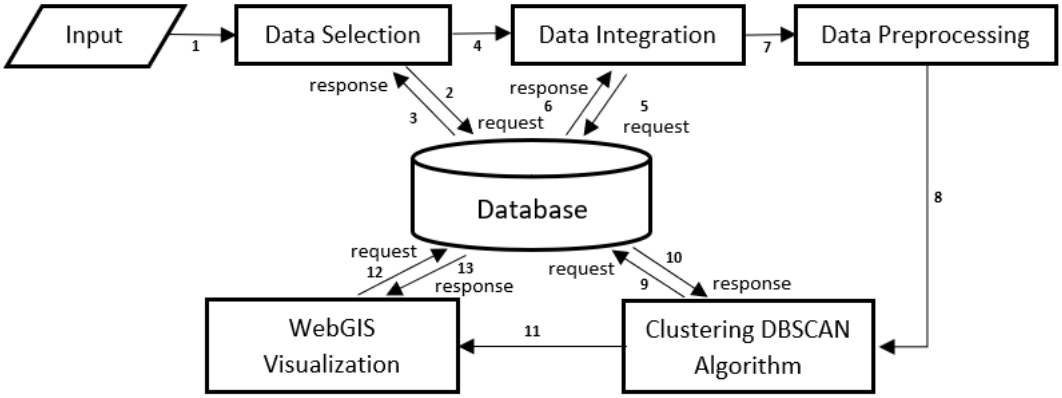


FIGURE 1. System design

Figure 1 shows our system design including data selection, data integration, data preprocessing, clustering, and visualization. Data selection is carried out by selecting features that are relevant to the analysis. The WebGIS-visualization that we created is a dynamic system, where the data in the database will always be enriched every time, the user carries out analysis with new data. This requires data integration, which is a process of combining two or more existing datasets into one dataset [9]. Data preprocessing includes data cleaning, data transformation, and dimensionality reduction. The data cleaning process is carried out to identify parts of data that are imperfect, incorrect, incomplete, inaccurate, or irrelevant [9]. Data transformation is used to change the data structure and attribute format [14]. In this process, the features will be converted into a form that can be accepted and processed by the DBSCAN algorithm. In data transformation, geocoding is also carried out to convert text-based addresses into geographic coordinates. Dimensionality reduction is used to compress data in such a way that the meaning of the data is not lost or simply reduces high-dimensional features to lower dimensions. In this study, dimensionality reduction was carried out through Principal Component Analysis (PCA) by identifying important features using the PCA concept. Next, the data is evaluated for performance improvement using the DBSCAN algorithm [18]. We integrate PCA and DBSCAN as follows: PCA performs formation of a new set of variables $N \times D$ (N is referred to as amount of data and D is referred to as Dimension), then evaluates the mean vector D and calculates the covariance. The final result of PCA is a new dataset which will be processed with the DBSCAN algorithm to form clusters. The PCA calculation is expressed in Equation (1).

$$PC_1 = a_{11} * X_1 + a_{12} * X_2 + \cdots + a_{1j} * X_j + \cdots + a_{1p} * X_p$$
$$PC_2 = a_{21} * X_1 + a_{22} * X_2 + \cdots + a_{2j} * X_j + \cdots + a_{2p} * X_p$$

$$PC_i = a_{i1} * X_1 + a_{i2} * X_2 + \cdots + a_{ij} * X_j + \cdots + a_{ip} * X_p$$
$$PC_p = a_{p1} * X_1 + a_{p2} * X_2 + \cdots + a_{pj} * X_j + \cdots + a_{pp} * X_p \quad (1)$$

In Equation (1), the a_{pj} are the coefficients associated with each original variable X_j in the p-th principal component PC_p. Figure 2 shows the visualization of 2D data after the PCA process.

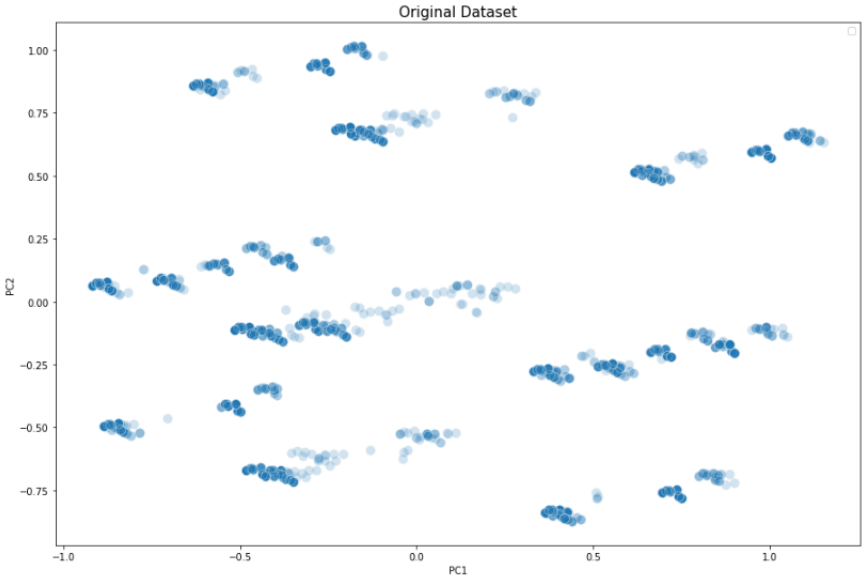
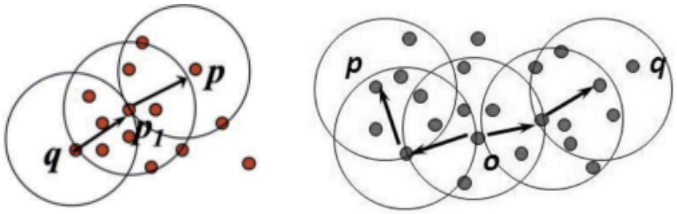


FIGURE 2. 2D visualization of dataset after PCA

After data preprocessing is complete, the DBSCAN algorithm is used for clustering. Compared with the K-Means clustering algorithm, which is only suitable for convex sample sets, the DBSCAN clustering algorithm can be applied to both convex and nonconvex sample sets [19]. DBSCAN is designed to find clusters and noise in a spatial database according to the density-reachability, density-connectivity, and identified noise, as shown in Figure 3.



(a) (b)

FIGURE 3. Illustration of density-reachable (a) and density-connected (b)

As shown in Figure 3(a), p is density-reachable from q, but q is not density-reachable from p. In Figure 3(b), p and q are density-connected to each other by o. Clustering with the DBSCAN algorithm uses two hyperparameters, namely epsilon (ε) and minimum point (*MinPts*). The ε is the maximum distance between a core point and other points. If data points are outside the ε range of the core point, then these points are considered outliers or noise. *MinPts* is the minimum requirement for data points that can form clusters. If data points in the ε range do not meet *MinPts*, then these points are considered outliers or noise.

There are 4 parameters used in forming clustering using the DBSCAN algorithm, namely neighborhood, core point, border point and outlier. Neighborhood is a term for data

points that are within the ε range of the central point and the amount of data meets *MinPts*. A core point is a point that is the center of the neighborhood ε. Border points are points which at the time of forming clustering do not meet *MinPts*, but one of the points in the neighborhood ε is part of another neighborhood ε. Outliers are a collection of data points that are merged into one neighborhood ε that does not meet *MinPts* and does not contain points that are core points in other neighborhoods [10].

MinPts is determined as follows. If the dataset is 2-dimensional, then *MinPts* is by default equal to 4. If the dataset dimension is more than 2, then *MinPts* is calculated using Equation (2), where D is the data dimension.

$$MinPts = 2 * D - 1 \tag{2}$$

The ε value can be determined using a heuristic approach, namely the ordered k-dist graph method which will measure the distance of each data point to other data points and will take the farthest distance between k nearest neighbors. In DBSCAN, k can refer to the *MinPts* value, because it can find the farthest distances that can meet the *MinPts* target. According to [20], all points with higher k-dist values (to the left of the threshold) are considered as noise, and all other points (to the right of the threshold) will be assigned to multiple clusters, as shown in Figure 4.

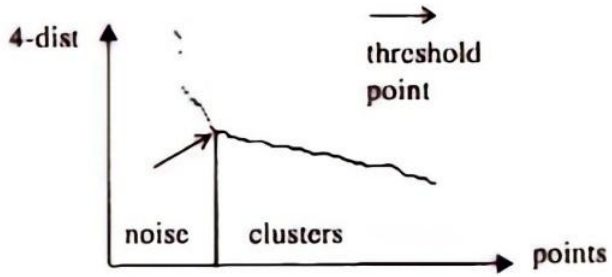


FIGURE 4. Ordered k-dist graph illustration

To find the optimal number of clusters, we perform hyperparameter tuning to select the optimal set of hyperparameters for the algorithm. To achieve this, we used the Sum of Squared Error (SSE) and Silhouette Score parameters. If the SSE value of a cluster is getting closer to 0, then the data points in each cluster have a very significant closeness. The SSE formula is expressed as Equation (3). The K value indicates the number of clusters, x_j defines the target object to each group C_i, and m_i is the reference point of each cluster.

$$SSE = \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - m_i\|^2 \tag{3}$$

Silhouettes (S) can be used to select optimal values from clusters with a suitable ratio scale to define clearly separated clusters. Clustering is considered as the average closeness between clusters that have similarities and differences [21]. The S score is expressed as Equation (4), where a defines the average distance between one data point and another data point in the same cluster, and b is the average distance between one data point and the data point in the next closest cluster.

$$S = \frac{b - a}{\{a, b\}} \tag{4}$$

3. **Result and Discussion.** Clustering was carried out using the DBSCAN algorithm which formed clusters based on two hyperparameters, namely ε and *MinPts*. The *MinPts* was determined using Equation (2). The dataset we used had dimensions of 27, resulting in a *MinPts* value of 53. The optimal ε value was determined using heuristic approach.

After carrying out several combination experiments between several ε values and *MinPts* of 53, the optimal ε value was obtained at 0.23. With a ε value of 0.23 and *MinPts* of 53, the DBSCAN algorithm produces 16 optimal clusters, as shown in Figure 5. The optimal number of clusters is determined based on the Goodness of Fit indicator which is shown by a small SSE value, and a Silhouette Score that is close to 1, as seen in Table 3.

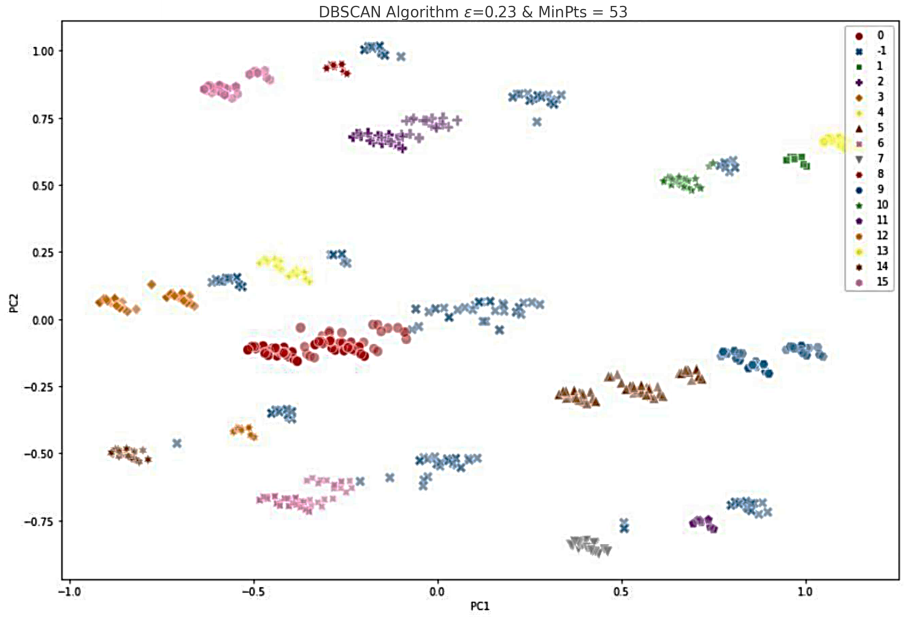


FIGURE 5. 2D visualization of formed cluster

TABLE 3. Clustering result

Algorithm	Number of clusters	SSE	Silhouette score
DBSCAN	16	71.22	0.805
K-Means	16	102.60	0.77

Table 3 shows the comparison of SSE and Silhouette Score between the DBSCAN algorithm and the K-Means algorithm. This comparison aims not only to validate the results, but also to find out how good these two algorithms are at forming clusters. The DBSCAN algorithm provides better results than the K-Means algorithm in forming clusters for $k = 16$ because it has a lower SSE and a higher Silhouette Score. DBSCAN has its own characteristics in identifying noise, so that the clusters formed can have a closer distance.

3.1. **Noise observation.** Table 4 shows that there are 278 noises in the formation of 16 clusters using the DBSCAN algorithm. DBSCAN defines noise as data points that are not close enough to the neighborhood center point. DBSCAN can detect noise in the input dataset, making it one of the advantages of DBSCAN because it can detect noise without performing this process in the data preprocessing stage. Noise can be data with erroneous values or attributes, or data that does not have a significant role (data that appears infrequently among other data).

3.2. **WebGIS visualization.** The clustering results are visualized in the form of geographic area mapping, namely mapping accident-prone locations, as shown in Figure 6. The dashboard will display a digital geographic map that maps traffic accident clusters to each coordinate location point. Cluster location points on the map have also been made more interactive by utilizing the tooltip feature which, when a location point is clicked, will display detailed cluster accident information at that location.

TABLE 4. Noise observation

No	Accident types	Vehicles involved	Noise	Note
1	OCA	PCV	30	Does not meet the *MinPts*
2	OCA	MCV-PCV	1	Attribute error
3	SVA	PCV	7	Does not meet the *MinPts*
4	CRA	PCV-MCV	8	Does not meet the *MinPts*
5	CRA	PCV-PCV	7	Does not meet the *MinPts*
6	CRA	TRV-PCV	2	Does not meet the *MinPts*
7	BSA	PCV	45	Does not meet the *MinPts*
8	BSA	PCV-PCV	1	Attribute error
9	BSA	MCV-PCV	1	Attribute error
10	HRA	PCV-PCV	48	Does not meet the *MinPts*
11	HRA	PCV-MCV	41	Does not meet the *MinPts*
12	HRA	PCV-Human	19	Does not meet the *MinPts*
13	HRA	PCV-PBV	1	Does not meet the *MinPts*
14	HRA	CTV-PCV	2	Does not meet the *MinPts*
15	HRA	TRV-PCV	8	Does not meet the *MinPts*
16	HRA	MRV-Human	3	Does not meet the *MinPts*
17	HRA	MRV-MRV	2	Does not meet the *MinPts*
18	HCA	TRV-TRV	4	Does not meet the *MinPts*
19	HCA	TRV-PCV	9	Does not meet the *MinPts*
20	HCA	TRV-Human	1	Attribute error
21	HCA	BsV-BsV	1	Does not meet the *MinPts*
22	HCA	MRV-PCV	2	Does not meet the *MinPts*
23	HCA	MRV-MRV	4	Does not meet the *MinPts*
24	HCA	MRV-BcV	1	Does not meet the *MinPts*
25	HCA	MRV-TRV	1	Does not meet the *MinPts*
26	HCA	BcV-PCV	2	Does not meet the *MinPts*
27	HCA	BcV-TRV	2	Does not meet the *MinPts*
28	HCA	PCV-PCV	21	Does not meet the *MinPts*
29	HCA	CTV-PCV	1	Does not meet the *MinPts*
30	MsA	MCV-PCV	1	Attribute error
31	MsA	TRV	1	Does not meet the *MinPts*
32	MsA	PCV	1	Does not meet the *MinPts*

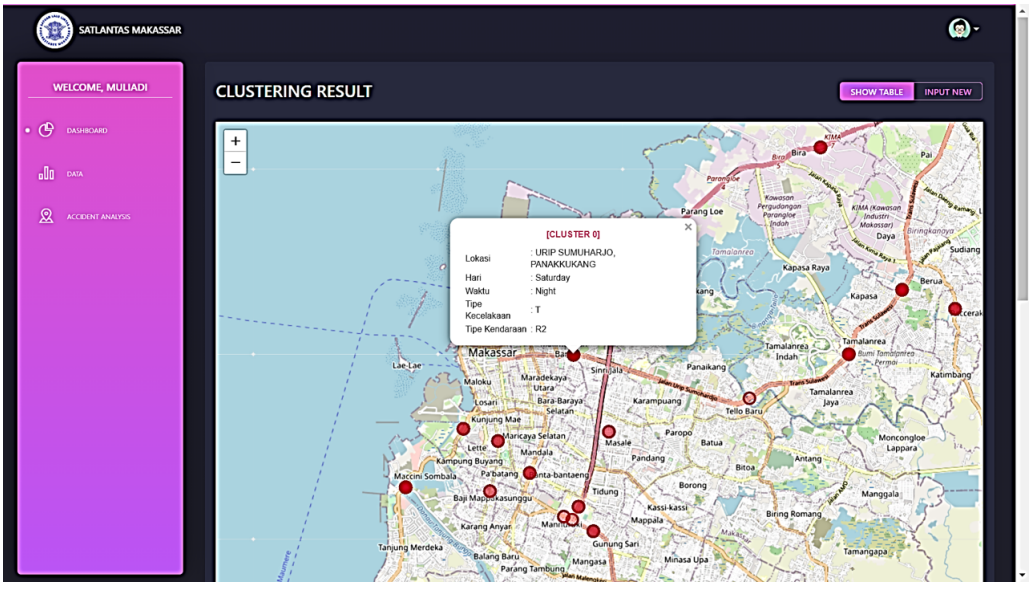


FIGURE 6. WebGIS visualization

4. **Conclusions.** Using a dataset consisting of 3558 daily traffic accident history data in Makassar City, Indonesia, the unsupervised DBSCAN algorithm produces 16 optimal clusters with an SSE value of 71.22 and a Silhouette Score of 0.805. This result is better when compared to the 16 clusters produced by the K-Means algorithm with an SSE value of 102.60 and a Silhouette Score of 0.77. The clusters produced by the DBSCAN algorithm have quite close intra-cluster cohesion and have quite far distances between clusters. In the future, the implementation of WebGIS visualization will be examined to speed up the execution time of the application.

Acknowledgment. This study was supported by the LBE grant, Faculty of Engineering, Universitas Hasanuddin, Indonesia.

REFERENCES

[1] WHO, *Road Traffic Injuries*, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, Retrieved on December 28, 2023.

[2] A. H. Adnan, H. Connie, H. Martha and P. Margaret, The political and social contexts of global road safety: Challenges for the next decade, *The Lancet Journal*, vol.400, no.10346, pp.127-136, 2022.

[3] S. Handayani et al., Safety culture in Indonesia's road transport, *Jurnal Management Transportasi dan Logistik*, vol.6, no.2, pp.201-210, DOI: 10.25292/j.mtl.v6i2.315, 2019.

[4] Y. Maqbool, A. Sethi and J. Singh, Road safety and road accidents: An insight, *International Journal of Information and Computing Science*, vol.6, 2019.

[5] M. R. Islam, I. J. Jenny, M. Nayon, M. R. Islam, M. Amiruzzaman and M. A. Al-Wadud, Clustering algorithms to analyze the road traffic crashes, *Proc. of International Conference on Science and Contemporary Technologies (ICSCT)*, Bangladesh, DOI: 10.1109/ICSCT53883.2021.9642542, 2021.

[6] D. Puspitasari, M. Wahyudi, M. Rizaldi, A. Nurhadi, K. Ramanda and Sumanto, K-Means algorithm for clustering the location of accident-prone on the highway, *International Conference on Advanced Information Scientific Development (ICAISD)*, West Java, Indonesia, DOI: 10.1088/1742-6596/1641/1/012086, 2020.

[7] K. Syahputri, R. M. Sari, I. Rizkya, U. Tarigan, I. Siregar and T. A. Farhan, Clustering the vulnerability of traffic accidents in Medan city with fuzzy C-Means algorithm, *The 2nd Talenta Conference on Engineering, Science and Technology*, Medan, Indonesia, DOI: 10.1088/1757-899X/801/1/012030, 2019.

[8] P. Dimitros and B. S. Ionnis, Data mining methods: A review, *International Journal of Computer Applicatins*, vol.183, no.48, 2022.

[9] X. Yuan, S. Gu, Z. Liu and L. Zhao, Mining multi-view clustering space with interpretable space search constraint, *IEEE Signal Processing Letters*, vol.30, 2023.

[10] P. Bhatia, *Data Mining and Data Warehousing: Principles and Practical Techniques*, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781108635592, 2019.

[11] A. G. Khosrowshahi et al., *Detecting Crash Hotspots Using Grid and Density-Based Spatial Clustering*, ICE Publishing, https://doi.org/10.1680/jtran.20.00028, 2021.

[12] X. Zhang, S. Su, H. Zhang, X. Zhang, Z. Chen and Y. Cui, Traffic accident location study based on AD-DBSCAN algorithm with adaptive parameters, *Proc. of the 26th International Conference on Computer Supported Cooperative Work in Design*, Brazil, DOI: 10.1109/CSCWD57460.2023.10152613, 2023.

[13] A. Gershtein and A. Terekhov, DBSCAN clustering method is applied to identify severe Traffic Accident (TA) hotpots on roads, *Computer Tools in Education*, vol.1, DOI: 10.32603/2071-2340-2020-46-58, 2021.

[14] A. P. Joshi and B. V. Patel, Data preprocessing: The techniques for preparing clean and quality data for data analytics process, *Oriental Journal of Computer Science and Technology*, vol.13, nos.2-3, pp.78-81, 2021.

[15] A. V. Vitianingsih, R. Marco, A. L. Maukar, E. W. Puspitarini and S. F. A. Wati, Spatial data modelling using MADM for classification of food self-sufficiency regions, *International Journal of Innovative Computing, Information and Control*, vol.18, no.2, pp.561-576, DOI: 10.24507/ijicic.18.02.561, 2022.

[16] D. R. Harris and C. Delcher, Bench4gis: Benchmarking privacy-aware geocoding with open big data, *IEEE Int. Conf. Big Data*, pp.4067-4070, DOI: 10.1109/BigData47090.2019.9006234, 2019.

[17] K. L. Khine and T. T. S. Nyunt, Predictive geospatial analytics using principal component regression, *International Journal of Electrical and Computer Engineering (IJECE)*, vol.10, no.3, pp.2651-2658, DOI: 10.11591/ijece.v10i3, 2020.

[18] S. Umadevi and N. S. Rajini, Dimensionality reduction of production data using PCA and DBSCAN techniques, *Intelligent Systems and Computer Technology*, vol.37, pp.458-462, DOI: 10.3233/APC200184, 2020.

[19] X. Tu et al., DBSCAN spatial clustering analysis of urban "production-living-ecological" space based on POI data: A case study of central urban Wuhan, China, *International Journal of Environmental Research and Public Health*, vol.19, https://doi.org/10.3390/ijerph19095153, 2022.

[20] J. Sander, M. Ester, H.-P. Kriegel et al., Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, *Data Mining and Knowledge Discovery*, vol.2, pp.169-194, https://doi.org/10.1023/A:1009745219419, 1998.

[21] T. Thinsungnoen, The clustering validity with silhouette and sum of squared errors, *Proc. of the 3rd International Conference on Industrial Application Engineering 2015*, 2015.