Journal of Education for Pure Science
Vol.16 , No.1 (2026)
DOI: https://doi.org/10.32792/jeps.v16i1.853

Website: jceps.utq.edu.iq *Email: jceps@eps.utq.edu.iq*

An Intelligent Model for Traffic Accident Hotspot Detection

Walaa Alajali [1,] **Wafaa Ali** [1] **Abdulrahman D. Alhusaynat**[2]

[1]Education College for Pure Sciences, University of Thi Qar,64001, Iraq

[2] Thi-Qar Education Directorate, Nassiriya, Thi-Qar, 64001, Iraq.
wafaa a@utq.edu.iq, rahmandakhil2@gmail.com

[*] Corresponding email: Walaakhshlan@utq.edu.iq

Received 1/ 10/2025, Accepted 17/ 12 /2025, Published 1 /3 /2026

Abstract:

Identifying traffic accident hotspots is essential for improving urban road safety and supporting Data-driven transportation planning. Despite the availability of extensive traffic data, many existing hotspot detection approaches are constrained by rigid spatial assumptions and limited integration of contextual risk factors. This study proposes a comprehensive framework for accident hotspot detection and risk characterization in Los Angeles, utilizing the 2023 U.S. accident dataset. After data cleaning and feature engineering, the data were analyzed using the Density-Based Spatial Clustering with Noise (DBSCAN) algorithm. The model identified 12 distinct hotspots. For each hotspot, a composite Danger Score was computed by combining normalized measures of accident frequency, severity, nighttime incidence, and adverse weather conditions. These scores were then partitioned into three risk levels: High (n = 4), Medium (n = 4), and Low (n = 4) using k-means clustering. Cluster validity was assessed using the Silhouette Score (0.565), Davies–Bouldin Index (0.439), and Calinski–Harabasz Index (2121.3), confirming a clear and robust spatial structure. The results show that reduced visibility, nighttime driving, and localized weather variability, particularly in high-density freeway and arterial segments, substantially contribute to increased accident risk. The proposed framework enables transportation agencies to proactively identify hazardous locations and prioritize targeted safety interventions in Los Angeles and other large metropolitan regions.

Keywords: Accident, DBSCAN, Traffic, Hotspot, Detection

1 Introduction

The challenges posed by road traffic accidents inrapidly growing urban areas are enormous and thuge social, economic, and infrastructural burden. Recent national data suggest that, compared to many non-urban areas, U.S. cities have an elevated crash risk due to growing mobility needs, exposure to multi-layered road networks, and disparate environmental factors. As a result, learning how to identify and describe areas of high concentrations of accident locations owing to the presence of particular factors, the number of crashes is higher than expected (hotspots), has become one of the main concerns on the part of transportation organizations wishing to improve traffic safety and optimize resource allocation.

Traditional classical hotspot detection approaches, such as kernel density estimation [1], spatial autocorrelation indices [2], and fixed-grid segmentation [3], have contributed to understanding crash distribution in space. However, these methods usually assume spatial structure in a rigid manner, are sensitive to parameter selection (e.g., bandwidth or grid size), and frequently under-value contextual risk factors such as visibility impairment, weather variability, or prevailing driving patterns over time. Further, a large proportion of studies focus on spatial frequency and ignore severity profiles and the interactions between environmental and temporal factors associated with collision risk. With the growing prevalence of real-time and historical traffic, weather, and infrastructure data becoming available to engineers, there is an opportunity to build flexible, data-driven analytics tools. Such models may be more successful in modeling the interactions between geographical, environmental, and temporal characteristics that contribute to incident risk within complex urban settings. Los Angeles is an especially useful case study, as it has one of the densest freeway systems, a large population, and continues to face traffic in various corridors. Although safety tasks are ongoing, the region still suffers from a high number of traffic accidents each year, which has implications for more sophisticated types of hotspot analysis methods. This dataset, U.S. Accidents 2023, provides a high-quality, large-scale collection of traffic accidents that includes a variety of geospatial, meteorological, and temporal attributes at a fine-grained level. This makes it ideal for detecting high-resolution hotspots and for risk profiling. Based on this collection, the current paper develops a DBSCAN-based hotspot detection framework for Los Angeles, along with a multivariate Danger Score to rate clusters by relative risk.

The current study fills certain gaps in the literature:

- Contextual risk synthesis: Several previous studies tend to focus on frequency rather than considering simultaneous combined effects, such as time, visibility, and weather, on accident risk.
- Strong cluster validation: The results of hotspot detection are commonly presented without structured analysis by several internal cluster validity indices.
- Actionable spatial risk stratification: Although spatial clusters are routinely detected, the process by which they are expressed as easily understandable, digestible risk categories that inform operational and policy decision-making is rare.

The novelty of this work are:

- A combined clustering analysis that applies DBSCAN to 2,972 accident records (2023), producing the most spatially coherent 12 hotspots and obvious noise.
- A composite multi-factor Danger Score is presented, which combines accident frequency, severity (number of victims), nighttime crashes, and bad weather situations to categorize high-risk, medium-risk, and low-risk hotspots.
- A robust validation architecture using Silhouette, Davies–Bouldin, and Calinski–Harabasz indices is implemented to assess the structural quality and separability of the detected clusters.
- An observational, temporal, and environmental study that dark driving and local weather are affecting the distribution of accidents in L.A.

The rest of the paper is structured as follows. We describe the dataset in Section 3. Section 4 describes the methodology, specifically feature engineering, DBSCAN clustering, and the Danger Score. Section 5 presents the findings for hotspot identification and risk classification. Section 6 concludes and presents implications for urban safety planning, as well as areas for future research.

2 Related Work

Traffic accident analysis research has evolved significantly over the years, where conventional statistical models such as multinomial logit, logistic regression, and a Bayesian approach were used to examine the associations between roadway, environmental, and human-related factors with crash outcomes [4], [5]. These methods yielded interpretable relationships but were based on strong distributional assumptions and had difficulties explaining the nonlinearities that are usually observed in accident datasets. Machine learning (ML) then offered a more flexible solution to these shortcomings. Methods, including decision tree, random forest (RF), support vector machine (SVM), and gradient boosting, have been utilized to crash data sets to improve prediction performance and deal with high-dimensional features [6]–[8]. Although ML models have displayed superior power to statistical-based models in complicated scenes, their ultimate performance also requires labour-intensive manual feature crafting and lacks the ability to model deeper spatial or temporal relationships between crash happenings. The advent of deep learning (DL) has greatly propelled the field. Convolutional Neural Network (CNN) can be a model of great potential to learn the spatial representations and extract latent feature structures from crash-related data [9]. Recurrent Neural Networks (RNNs) and more specifically Long Short-Term Memory networks (LSTMs) have also been used to model temporal dependencies between historic crashes [10]. Hybrid CNN and LSTM architectures further allow simultaneous inference of spatial and temporal patterns, and have demonstrated better results in reporting crash severity [11].

Even more advanced techniques added attention mechanisms to increase the interpretability of predictive models and the ability to focus on key aspects; this is especially pertinent when dealing with imbalanced data, where severe crash data is underrepresented [12], [13]. Furthermore, in reducing redundancy and noise within high-dimensional crash datasets, dimensionality reduction methods such as Principal Component Analysis (PCA) have been implemented [14]. In recent studies, multi-view and hybrid deep learning approaches have been explored to consider nonlinear relationships between variables affecting crash severity [15]. Nevertheless, most of the existing studies mainly focus on predicting crash severity rather than hotspot locations. Few studies incorporate spatial clustering with environmental and temporal contextual factors, such as weather conditions, visibility, or nighttime patterns, which are key to understanding why some locations become chronic high-risk areas. This introduces a disconnect between crash-level severity modeling and network-wide road safety analysis, where predicting hotspots is important towards proactive urban planning for safety.

3 Dataset Description

We use the U.S. Accidents Dataset (Version 2023) [16], which is a comprehensive resource of traffic accidents captured from various disparate sources over the United States. This dataset contains a comprehensive collision database, which is enriched with data from traffic cameras, intelligent transportation system (ITS) sensors, and navigation apps, as well as state DOTs as shown in Figure 1. The 2023 update comprises over 130K validated accident reports containing geolocation and broad contextual variables.

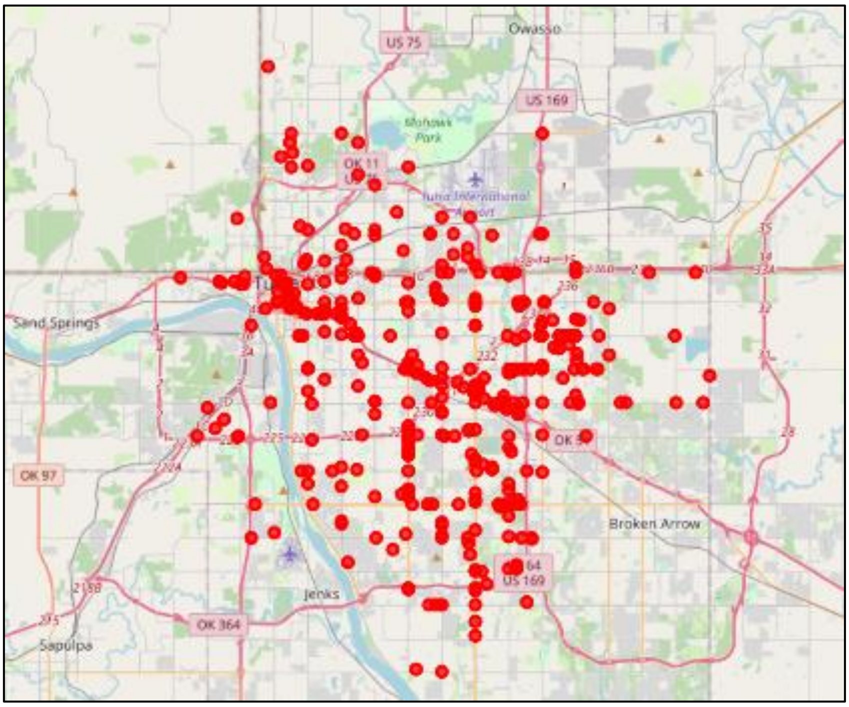

Figure 1: Map of Los Angeles Accidents Locations

3.1 Data Collection and the Study Area

The raw data were sourced from a public open data project related to U.S. traffic accidents. The original dataset included accidents from roughly 49 states, excluding areas with poor reporting mechanisms. On this occasion, the data was filtered to only include accidents with their city field equal to "Los Angeles" and where the accident date is within 2023. Upon filtering and preprocessing, 2,972 geolocated accident entries were left. All these data are with spatial-temporal points, as well as various factors of environment and roadways information, which can be used for the analysis of urban accident patterns at fine resolution.

Los Angeles is crisscrossed by numerous major interstates and U.S. highways (e.g., I 5, I 10, US 101, I 405) as well as arterial streets and local streets. These features, coupled with heavy traffic and frequent congestion, make the city a typical and challenging case for studying accident hotspots.

4 Methodology

The integrated analytical model applied in this research consists of both spatial clustering and contextual feature engineering, as well as composite risk quantification to perform a full coverage accident hotspot analysis for Los Angeles. The workflow is divided into four main steps: (1) data preparation, (2) feature engineering, (3) unassisted spatial clustering with DBSCAN, (4) risk level classification by means of a composite Danger Score.

4.1 Analytical Steps:

The entire process is based on a stepwise data-driven approach:

1. Load and clean accidents records data set.
2. Build a time span and environmental characteristics that give the context of each occurrence.
3. Perform DBSCAN clustering in geographical space to recognize spatially dense accident locations (hotspots) and noise points.

4. Calculate the combined Danger Score for each cluster considering Frequency, Severity, Nighttime share, and Adverse weather share.
5. Group clusters by danger level according to k-means of Danger Scores and measure quality of clustering with a variety of internal indices.

4.2 Data Preprocessing

4.2.1 Spatial and Temporal Filtering

The accident records were kept if and only if they: (a) had an accident date in 2023; | City="Los Angeles" & 0 = Lat >= 200|; ave prev solution | valid Latitude= &&370 Longitude within the metropolitan area bounds. Time fields were normalized to the datetime type for consistent extraction of hour, day, and month features.

4.2.2 Handling Missing Data

Continuous weather variables (visibility, precipitation, wind speed, pressurehumidity) had few missing values (i.e., not more than 0.01% of the cases) with imputation based on median values computed over the Los Angeles sample subset. This approach retains centrality without assuming normality and is robust to outliers. Any record with a null or invalid spatial or time field were excluded from the original dataset because they could cause the defensibility of geographical clustering to fail.

4.2.3 Outlier Treatment

Spatial outliers, defined as coordinates well outside the Los Angeles's area for this analysis, were removed in a pre-processing process. Any rarely showing isolated accidents had not been filtered out: DBSCAN was still allowed to label such points as noise in the clustering process.

4.3 Feature Engineering

4.3.1 Temporal Features

To consider temporal changes in the patterns of accidents, we derived the following features:

- Hour (0–23)
- Month (1–12)
- DayOfWeek (0–6)
- IsNight, which represents slip accidents from 7 pm to 6 am

These can aid in investigations of daily, weekly, and seasonal trends in crash rates or severity.

4.3.2 Meteorological and Environmental Features

Weather conditions were characterized by means of continuous and binary indicators:

Continuous: Visibility, Precipitation, Wind_Seed, Humidity, and Pressure as shown in Figures [2-6]

A binary indicator, AdverseWeather, that indicates heavy rain, fog, storm, or any other such thing that may make the road condition worse, derived from Weather_Condition.

Such properties permit the study to explore implications of localized differences in weather and visibility in accident risk.

4.3.3 Severity Encoding

Variable severity (1–4 as in the dataset) was kept in its original ordinal form. This maintains the ordered nature of severity outcomes and simplifies the computation of CLMS.

4.4 Spatial Clustering with DBSCAN

4.4.1 Rationale for DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) was chosen as it:

- Identifies arbitrarily-shaped clusters in geographic space.
- Do not need to specify the number of cluster in advance.
- It explicitly delineates noise, which helps distinguish one-off accidents from repeated hotspots.
- Is resistant to outliers and robust for urban realistic datasets with disparities in densities between the locations.

These characteristics make DBSCAN appropriate to model urban accident patterns and hotspots of interest.

4.4.2 Choice of Distance Metric and Parameters

Clustering was done with the Haversine distance metric on latitude and longitude converted into radians to have geodesic correct distance on Earth's surface. DBSCAN requires two main parameters:

eps: the maximum neighborhood radius, min_samples (MinPts): the minimum number of samples in a region. Non-tree matching was performed by searching a grid of (eps, MinPts) values and selecting the candidate configurations based on a Silhouette Score. Example results include:

eps = 0.03, MinPts = 15 → Silhouette=0.8074

eps = 0.03, MinPts = 25 → Silhouette = 0.9319

eps = 0.05, MinPts = 25 → Silhouette = 0.7612

The setting with eps = 0.03 and MinPts = 25 obtained the best Silhouette Score (0.9319) in the tuning phase, so it has been selected as the optimal parameter set. Ultimately, DBSCAN revealed 12 clusters and 497 noise points (16.72% of total accidents), which is reflective of the existence of dense urban hotspots as well as dispersed incidents in Los Angeles.

Score Components

For each cluster , we compute four components:

- Accident frequency: total of accidents related to all accidents in clusters.
- Average severity: the average (mean) severity on the 1–4 scale.
- Nighttime proportion: percentage of accidents in the cluster involving nighttime.
- Adverse weather proportion: ratio of accidents included that happened under adverse weather.
- Both components range from 0 1, so I then combined them: The Danger Score =2.
- The weights were subjectively tuned to place more weight on accident frequency and accident severity, but also consider environmental conditions.

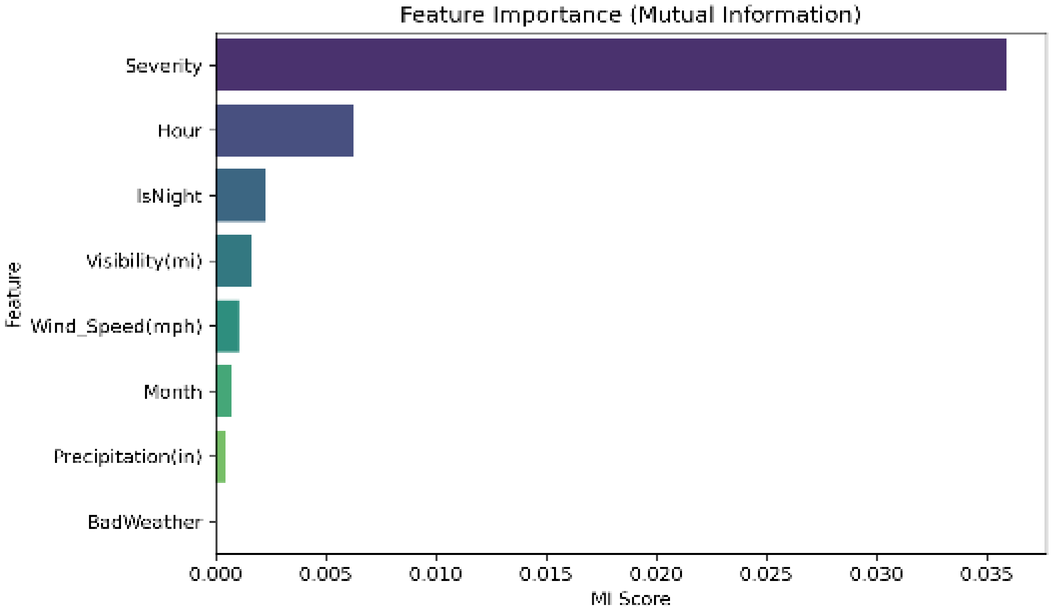

Figure 2: Feature Importances

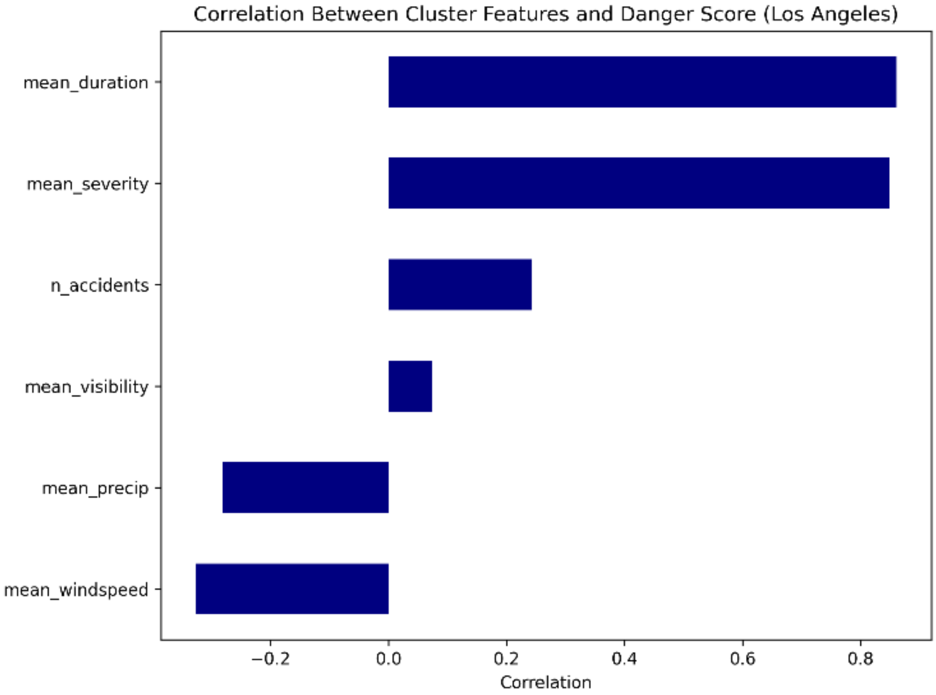

Figure 3: Feature Correlation

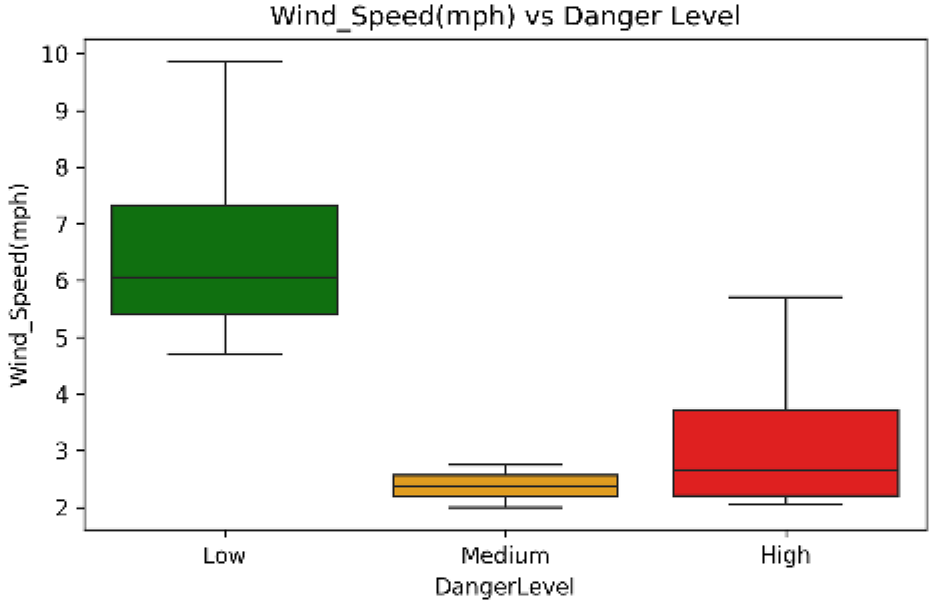

Figure 7: The correction between Danger Level and Wind Speed

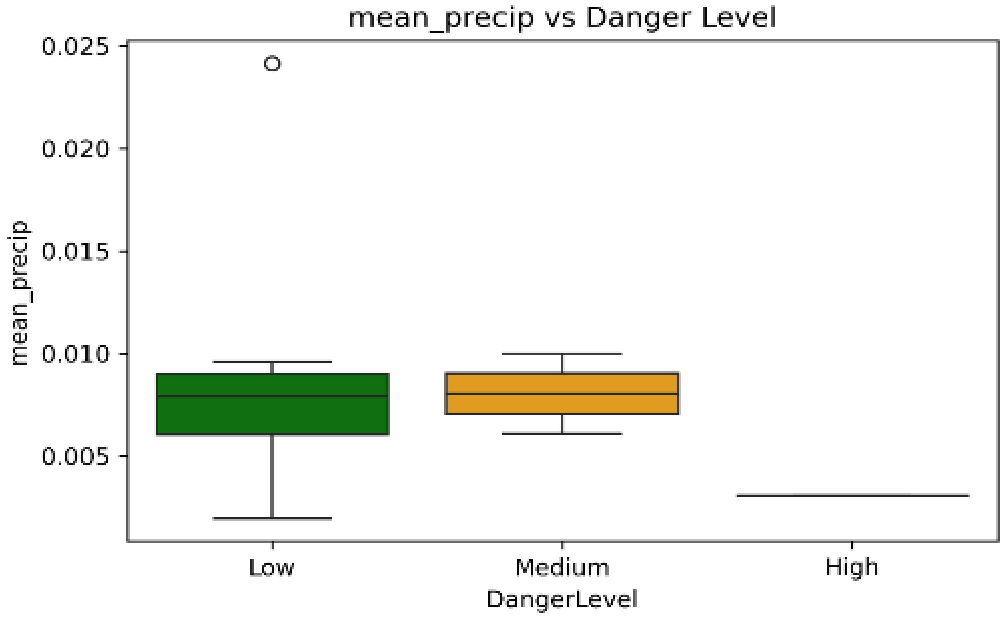

Figure 4: The correction between Danger Level and Precipitation

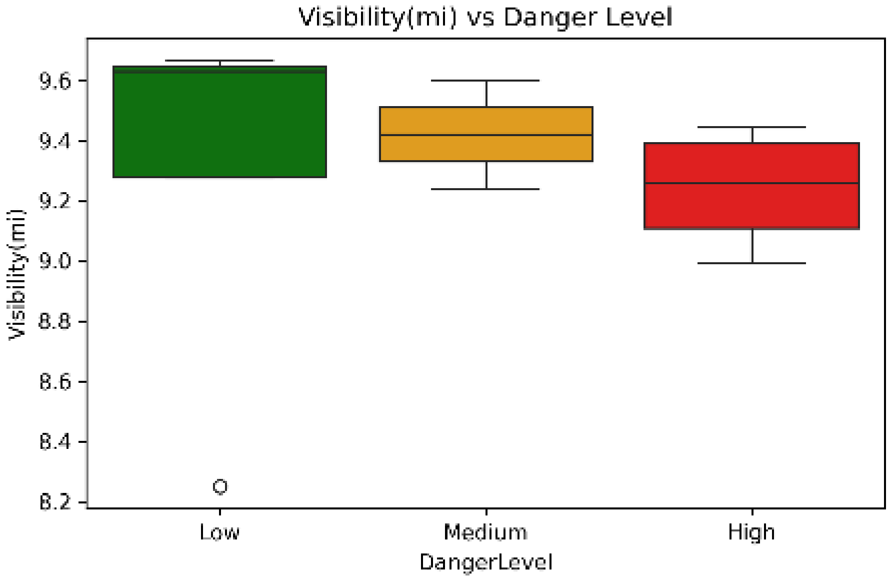

Figure 5: The correction between Danger Level and Visibility

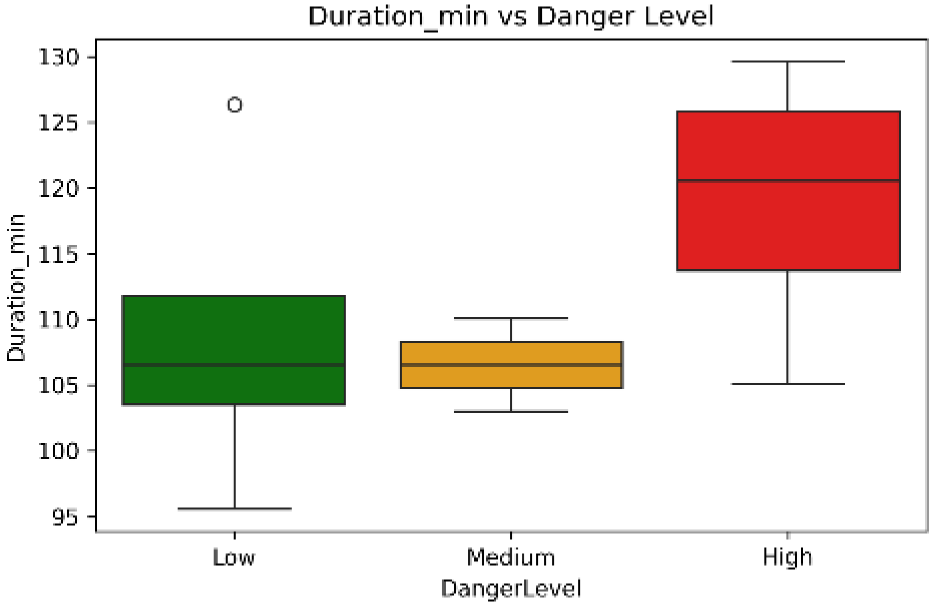

Figure 6: The correction between Danger Level and Duration

5 Results

5.1 Spatial Hotspot Identification

By utilizing the selected DBSCAN parameter (eps = 0.03, MinPts =25), 12 spatially isolated hotspots were identified in Los Angeles alongside with 497 noise points, accounting for a total of accidents. These hotspots are predominantly located near freeway interchanges and off ramp access, along heavily trafficked arterial corridors, on sites of repeated congestion and local obstructed weather or visibility. Cluster sizes are widely variable (e.g., small local hot-spots, large high-density regions that abut major transport infrastructure).

5.2 Cluster-Level Accident Characteristics

The cluster-level profiles exhibited wide variation in accident attributes such as frequency, severity distribution, and situational context. Largest hotspot included about 1,600 crashes and smaller hotspots had from 25 to 120 crashes, which showed a significant variety of spatial concentration. In addition to frequency, emission clusters were found to have a significantly higher proportion of severity level 3–4 crashes at some locations where simpler crash events are relatively less frequent. There were also marked differences in contextual factors between clusters. In some of these hotspots, the values of visibility reduction, humidity, and precipitation were higher than average, indicating that local climatic factors contribute to a higher vulnerability in those areas. Those differences underscore that the accidents are not distributed evenly throughout the road network, but are more dependent on cluster-specific temporal behavior and weather exposure.

5.3 Distribution of Danger Scores

The mixed Danger Score over all the clusters effectively confirms a born hierarchy of risk as shown in Figure 2.

-High-risk cluster (n = 4): High accident frequency, high average severity and disproportionate night-time and bad-weather accidents.

-Medium-risk clusters (4): Average incidents in quantity and overall mix severity profile, some of the background and temporal risk factors might show up, but not quite to a very extreme.

-Low-risk clusters (n = 4): The cluster groups that contain less number of accidents, or with fewer severity patterns and low adverse weather / nighttime effect.

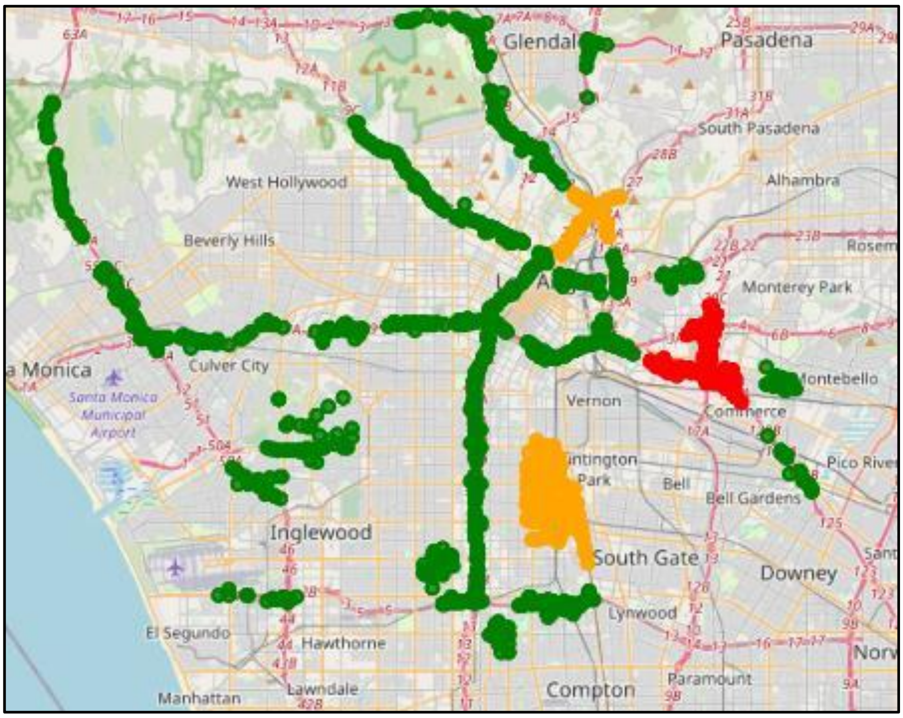

Figure 2: Accident Hotspot Clusters

5.4 Clustering Validation

The stability of the clustering results is also assessed with internal validation indices:

Silhouette Score=0.565 :- Good distance between the clusters.

Davies–Bouldin Index = 0.439: Representative of dense clusters, with a small ratio of between-cluster distance to inter-cluster similarity.

Calinski–Harabasz Index = 2121.3: Indicates that the between-cluster dispersion is much larger than the within-cluster dispersion.

All these indices suggest that DBSCAN can capture the stable and significant hotspot patterns for use in urban safety analyses.

6 Conclusion

In this study, crash risk is influenced by a mix of built-environment factors, temporal exposures, and environmental conditions. At the cluster level, significant heterogeneity of accident frequency, severity, and situation was observed, indicating that traffic crashes are not evenly distributed over the urban network. Night driving in particular appeared as a common contributor, especially when sections have, at the same time, complex road geometry, higher actual traffic speed, or insufficient illumination. Environmentals such as mist, haze, humidity, and wet pavement were highly significantly correlated with high DangerScores, indicating the important role of visibility and weather effects on collision probabilities. These effects were most evident in semi-arid locations where infrequent rain events minimize the opportunities for drivers and infrastructure to become conditioned to wet conditions.

Risk was also differentiated by characteristics of the infrastructure. High- and medium-risk clusters were often situated near on-ramps, off-ramps, and intersections with high volumes of traffic--features that inherently create closely spaced conflict points which rear-end crashes and side-impact collisions thrive. On the contrary, low-risk clusters were more likely to play out along simplified corridors with less navigational pressures and lower exposure levels. For our future direction, development of real-time accident monitoring and risk predication could be considered for proactive safety control and dynamic decision support.

References

[1] **J. Yang, Q. Liu, and M. Deng**, “Spatial hotspot detection in the presence of global spatial autocorrelation,” *International Journal of Geographical Information Science*, vol. 37, no. 8, pp. 1787–1817, 2023, https://doi.org/10.1080/13658816.2023.2219288

[2] **L. Thakali, T. J. Kwon, and L. Fu**, “Identification of crash hotspots using kernel density estimation and kriging methods: A comparison,” *Journal of Modern Transportation*, vol. 23, pp. 93–106, 2015, https://doi.org/10.1007/s40534-015-0068-0

[3] **X. Wang, Z. Zhang, and Y. Luo**, “Clustering methods based on stay points and grid density for hotspot detection,” *ISPRS International Journal of Geo-Information*, vol. 11, no. 3, p. 190, 2022, https://doi.org/10.3390/ijgi11030190

[4] **Alsaleh, R., Walia, K., Moshiri, G., & Alsaleh, Y. T**. (2025). Traffic Collision Severity Modeling Using Multi-Level Multinomial Logistic Regression Model. *Applied Sciences*, *15*(2), 838. https://doi.org/10.3390/app15020838

[5] **Ali Kemal Çelik, Erkan Oktay**, “A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey”, Accident Analysis & Prevention, Volume 72, 2014, Pages 66-77, https://doi.org/10.1016/j.aap.2014.06.010.

[6] **Ittirit Mohamad, Sajjakaj JomnonKwao, Vatanavongs Ratanavaraha**, “Machine learning predictive performance in road accident severity: A case study from Thailand”, Results in Engineering, Volume 26, 2025, 104833, https://doi.org/10.1016/j.rineng.2025.104833.

[7] **Elyassami, S., Hamid, Y., & Habuza, T**. (2021, June). “Road crashes analysis and prediction using gradient boosted and random forest trees”. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)* (pp. 520-525).

[8] **Zhibin Li, Pan Liu, Wei Wang, Chengcheng Xu**, Using support vector machine models for crash injury severity analysis, Accident Analysis & Prevention, Volume 45, 2012, Pages 478-486, https://doi.org/10.1016/j.aap.2011.08.016.

[9] **Jia Li**, "Deep learning model for predicting traffic accident severity under imbalanced data," Proc. SPIE 13575, International Conference on Smart Transportation and City Engineering (STCE 2024), 135753A (28 April 2025); https://doi.org/10.1117/12.3062104

[10] **Leinonen, M., Al-Tachmeesschi, A., Turkmen, B.** ***et al***. Long Short Term Memory Based Traffic Prediction Using Multi-Source Data. *Int. J. ITS Res.* **23**, 354–371 (2025). https://doi.org/10.1007/s13177-024-00451-y

[11**] Zhao, J., Yang, W., & Zhu, F.** (2024). A CNN-LSTM-Attention Model for Near-Crash Event Identification on Mountainous Roads. *Applied Sciences*, *14*(11), 4934. https://doi.org/10.3390/app14114934

[12] **G. Rajesh, A. R. Benny, A. Harikrishnan, J. Jacob Abraham and N. P. John**, "A Deep Learning based Accident Detection System," *2020 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, 2020, pp. 1322-1325, doi: https://doi.org/10.1109/ICCSP48568.2020.9182224

[13] **F. Liu, H. Wang, and Y. Chen**, "Explainable deep attention networks for traffic crash severity analysis," *Accident Analysis & Prevention*, vol. 171, 2022, https://doi.org/10.1016/j.aap.2022.106670

[14] **Achroo, P., Yadav, A., Kathuria, A., Agarwal, S., Islam, M.M.** (2024)." Dimensionality Reduction and Machine Learning-Based Crash Severity Prediction Using Surrogate Safety Measures". 2022. Lecture Notes in Civil Engineering, vol 443. Springer, Singapore. https://doi.org/10.1007/978-981-99-7976-9_56

[15 **Trirat, P., Yoon, S., & Lee, J. G.** (2023). MG-TAR: Multi-view graph convolutional networks for traffic accident risk prediction. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 3779-3794.

[16] **S. Moosavi**, "U.S. Accidents (2016–2023) Dataset," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents.