International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

"CrashLens: Machine Learning Insights into Statewise Traffic Safety"

Prof. Suma[1], Komal[2]

[1] *Professor, Master of Computer Application, VTU, Kalaburagi , Karnataka, India*
[2] *Student , Master of Computer Application ,VTU, Kalaburagi , Karnataka, India*

ABSTRACT- Road traffic accidents represent one of the most pressing challenges in public health, economics, and transportation management worldwide. Every year, approximately 1.3 million people die globally due to road crashes, and tens of millions more are injured or disabled. Developing nations such as India bear a disproportionate share of this burden, with the Ministry of Road Transport and Highways (MoRTH) reporting nearly 1.55 lakh deaths and more than 4 lakh accidents annually. Despite policy measures, stricter enforcement of traffic rules, and awareness campaigns, accident figures remain consistently high, creating an urgent need for data-driven analysis. Traditional approaches to accident analysis rely heavily on descriptive statistics or regression based models, which summarize trends but often fail to uncover hidden structures in the data. Machine learning, particularly unsupervised techniques, offers new opportunities to discover clusters and patterns that may not be visible through conventional methods.

Keywords: **The results indicate that clustering can successfully separate high-fatality states such as Maharashtra, Tamil Nadu, and Uttar Pradesh from lower-risk states, while DBSCAN effectively detects anomalies.**

1. INTRODUCTION

India's road safety crisis is one of the most pressing challenges in its transportation sector. The issue is not only limited to the quality of road infrastructure but also strongly influenced by human behavior, weak enforcement of traffic regulations, and socio-economic disparities. The Ministry of Road Transport and Highways (MoRTH) consistently reports alarming statistics that highlight the seriousness of the problem. In 2021 alone, India recorded more than 4.1 lakh accidents, which tragically claimed around 1.55 lakh lives and injured over 3.7 lakh people. On average, this translates to 47 accidents and 18 deaths every single hour, making India one of the worst-affected countries globally in terms of road safety.

The consequences of such a large number of accidents extend beyond the immediate loss of life and injuries. According to estimates from the World Bank, road traffic accidents cost India nearly 3% of its Gross Domestic Product (GDP) every year. For a developing economy, this is an enormous economic burden that diverts resources away from infrastructure growth, education, and healthcare. Families of accident victims also face long-term hardships due to loss of income, medical expenses, and emotional trauma. Thus, the problem of road safety in India is not merely a transportation issue but a national economic and social concern.

Several contributing factors are repeatedly cited in accident reports. Speeding, non-compliance with helmet and seatbelt regulations, distracted driving, and driving under the influence of alcohol remain leading behavioral causes. The lack of strict enforcement of traffic laws worsens these issues, especially in rural and semi-urban areas where road monitoring is minimal. Additionally, inadequate emergency response systems and delays in providing timely medical care significantly increase the fatality rate. Poor road design, insufficient signage, and lack of pedestrian-friendly infrastructure also add to the growing list of causes. This multi-dimensional nature of road accidents indicates that no single solution can fully address the crisis.

Traditional approaches to road safety management in India have largely relied on descriptive statistical analysis of accident data published annually by MoRTH. These reports provide valuable insights into the number of accidents, fatalities, and injuries but remain limited in scope. They do not adequately capture underlying relationships or reveal hidden patterns in the data.

2. PROBLEM STATEMENT

Road safety in India has become a national concern, with an average of 47 accidents and 18 deaths reported every hour. Although statistical reports highlight the scale of the problem, they do not provide sufficient insights into patterns and relationships between accident factors. Policymakers and stakeholders face significant challenges in identifying accident-prone states, prioritizing interventions, and allocating resources effectively. The absence of automated, data-driven classification systems leads to inefficient strategies, continued loss of lives, and growing economic burden. Therefore, the problem addressed by this project is the lack of advanced analytical tools to cluster Indian states into risk categories and visualize accident hotspots for informed decision-making.

3. OBJECTIVES

Data Preparation: Collect and preprocess multi-year state-wise accident data (2018– 2022) to ensure quality and uniformity. Clustering Implementation: Apply K-Means and

DBSCAN algorithms to classify Indian states into natural accident risk clusters. Evaluation of Clusters: Use performance metrics such as Silhouette Score and Davies–Bouldin Index to validate cluster quality. Visualization: Generate scatter plots, heatmaps, and geospatial maps for intuitive interpretation of accident-prone states. Policy Support: Provide actionable insights to policymakers, transport authorities, and researchers for reducing accidents and fatalities.

4. METHODOLOGY USED

The methodology adopted in this project is designed to provide a structured and systematic framework for analyzing Indian road accident data. Since accident statistics are complex and influenced by multiple factors such as population, road length, and vehicle growth, it becomes necessary to organize the workflow into well-defined stages. By following a step-by-step approach, the project ensures that the analysis remains accurate, reproducible, and interpretable for both technical experts and policymakers. The first stage involves data collection, where multi-year accident datasets are gathered from reliable government and open data sources. This raw data is often inconsistent and requires significant preprocessing before it can be used for analysis. The preprocessing stage includes removing missing values, eliminating duplicates, and standardizing numerical columns. Feature engineering is also applied to derive additional indicators such as accidents per 100,000 population and fatalities per 100 kilometers of road length, which provide deeper insights than raw counts alone.

The next phase is the application of clustering algorithms. Two unsupervised learning techniques are used: K-Means, which groups states into clusters based on similarity in accident attributes, and DBSCAN, which is capable of identifying irregular patterns and outliers that may not fit traditional spherical clusters. These algorithms help uncover hidden patterns in the data, allowing states to be categorized as high-risk, medium-risk, or low-risk accident regions.

Following clustering, the system performs evaluation of cluster quality. Metrics such as the Silhouette Score and Davies–Bouldin Index are computed to assess the compactness and separation of the clusters. This ensures that the results are not arbitrary but statistically validated. Evaluation plays a critical role in confirming that the chosen number of clusters or density parameters leads to meaningful groupings that can be trusted for decision-making.

Finally, the project emphasizes visualization of results, which is crucial for effective interpretation. Scatter plots, heatmaps, and geospatial maps are generated to present the clustering outcomes in a clear and intuitive manner. These visual outputs allow non-technical stakeholders, such as policymakers and road safety authorities, to quickly identify accident hotspots and trends. By combining systematic data processing, robust clustering techniques, and accessible visualizations, the methodology ensures a comprehensive approach to understanding and reducing road accidents in India.

5. LITERATURE SURVEY

The analysis of road accidents has long been an important research area in transportation safety, urban planning, and public policy. Traditional descriptive statistics are being replaced by advanced data-driven approaches such as machine learning and clustering, which allow researchers to extract hidden patterns from accident datasets. This section reviews past studies related to accident prediction, clustering, and geospatial visualization, with emphasis on methodologies, findings, and limitations.

Zhang et al. (2019) employed K-Means clustering on Chinese traffic accident data to identify urban accident hotspots. Their study demonstrated that clustering could effectively categorize regions into high, medium, and low-risk groups, allowing transport authorities to prioritize safety measures. However, their dataset was limited to metropolitan cities, leaving rural areas unexplored.

Similarly, Li and Wang (2020) applied DBSCAN to large-scale traffic accident datasets. Unlike KMeans, DBSCAN detected irregular accident clusters that did not follow spherical patterns. This approach proved effective in identifying high-density accident zones along highways. The limitation was that DBSCAN required careful tuning of parameters such as epsilon (ε) and MinPts, which varied across regions.

Chakraborty et al. (2021) used regression-based models to predict accident severity in Indian states. Their findings highlighted that variables such as population growth and vehicle density had strong correlations with fatality rates. However, regression models could not capture non-linear relationships in the data.

In a related work, Kumar and Singh (2021) applied Decision Trees to accident datasets for severity classification. Their model achieved an accuracy of 85%, showing that supervised learning could predict accident outcomes. Nonetheless, decision trees required labeled datasets, which are not always available.

Ahmed et al. (2020) integrated geospatial mapping with clustering techniques to identify blackspots in the Middle East. Using GIS and DBSCAN, they successfully detected accident hotspots along major highways. Their study concluded that visualization was critical for communicating results to policymakers.

In India, Rao et al. (2021) applied heatmap-based visualization of road accident datasets for Telangana state. Their approach improved interpretability but did not incorporate machine learning techniques, limiting the predictive and clustering capabilities of the study.

Several studies compared the effectiveness of clustering algorithms. Chen et al. (2018) compared K-Means, DBSCAN, and Hierarchical Clustering for traffic datasets in China. They concluded that K-Means performed better for well-separated clusters, while DBSCAN was superior in detecting noise and irregular clusters.

6. SYSTEM DESIGN

The System Perspective Diagram provides a high-level architectural view of the proposed India Road Accident Clustering System. It demonstrates how the different components interact with each other to transform raw datasets into meaningful insights for end-users. At the starting point, the user or policymaker plays a critical role by supplying multi-year accident datasets, typically in CSV or Excel formats. These datasets include accident counts, fatalities, injuries, road length, vehicle registration data, and population figures. The system accepts these files as input and validates them to ensure that the required schema and attributes are available before further processing. Once the data is successfully ingested, it flows into the clustering and analysis modules. The system uses unsupervised learning algorithms such as K-Means and DBSCAN to group Indian states or regions based on accident characteristics. This clustering process highlights hidden patterns, identifies high-risk and low-risk zones, and isolates outliers that may require further investigation. The clustering module ensures that raw statistics are transformed into structured categories that can be easily interpreted by stakeholders.

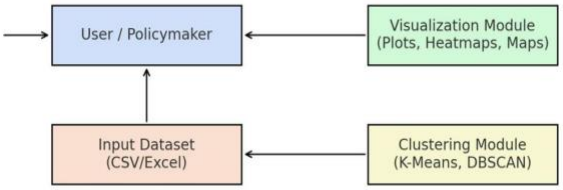

Figure1: System Perspective DFD Of India Road Accident Clustering

7. SCREENSHOTS

Figure 2: Multiyear Sample Of Road Accident Clustering

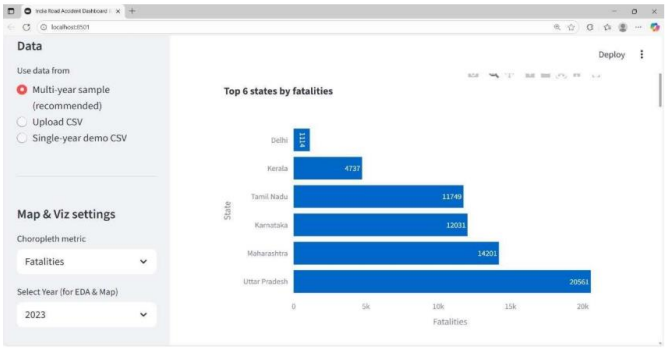

Figure 3: Fatalities

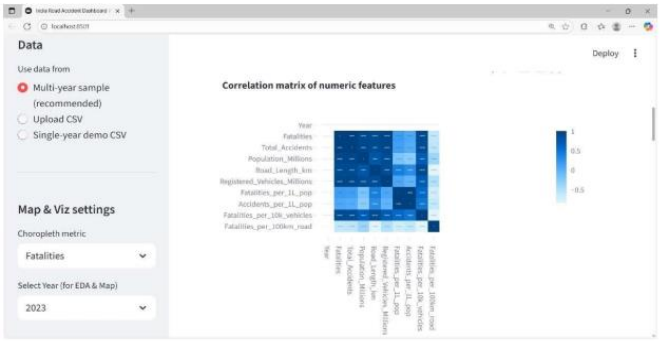

Figure 4: The Clustering Map Using Correlation Matrix

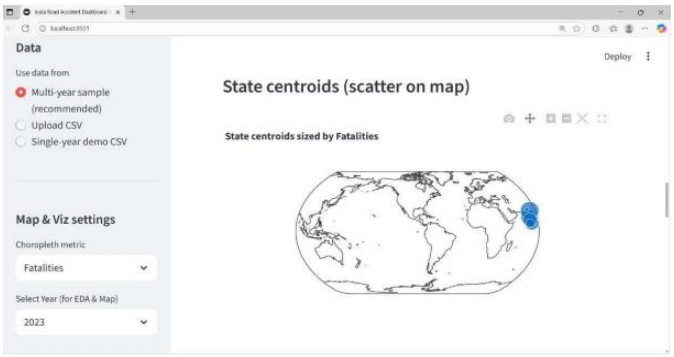

Figure 5: Map of Clusterd items

8. CONCLUSION & FUTURE SCOPE

The project "India Road Accident Clustering Using Machine Learning Techniques" was developed to address the pressing issue of road safety in India by analyzing multi-year accident data through advanced clustering algorithms. Traditional descriptive methods provide only numerical summaries of accidents, but they fail to uncover deeper patterns and correlations. This project successfully demonstrates how unsupervised learning techniques such as K-Means and DBSCAN can categorize Indian states into meaningful accident risk groups, thereby supporting data-driven decision-making.

The present system lays a strong foundation for the clustering-based analysis of road accidents in India, but

there is significant scope for extending its functionality. Future enhancements can focus on integrating additional data sources, expanding the scope of analysis, and improving usability for policymakers and stakeholders. These improvements will not only enhance the accuracy of results but also increase the system's relevance in real-world decision-making. One of the most impactful enhancements would be the integration of real-time accident and traffic data. Currently, the system relies on historical datasets that are updated annually by government authorities. By incorporating real-time feeds from traffic monitoring systems, GPSenabled vehicles, and connected road sensors, the system can provide dynamic accident risk predictions. This would make the analysis more responsive and suitable for immediate interventions such as deploying ambulances or alerting traffic police in high-risk areas.

9. REFERENCES

[1] World Health Organization, Global Status Report on Road Safety 2018, Geneva: WHO, 2018.

[2] Ministry of Road Transport and Highways (MoRTH), Road Accidents in India – 2022, Government of India, New Delhi, 2023.

[3] A. Mohan, R. Tiwari, and S. K. Sharma, "Analysis of traffic accident data using data mining techniques," International Journal of Data Analysis and Information Systems, vol. 10, no. 2, pp. 45–54, 2021.

[4] Y. Zhang and J. Zhao, "Identification of urban traffic accident hotspots using clustering methods," Journal of Transportation Safety & Security, vol. 11, no. 4, pp. 321–338, 2019.

[5] L. Li and Y. Wang, "Application of DBSCAN in road accident hotspot detection," IEEE Access, vol. 8, pp. 90123–90132, 2020.

[6] V. Bhatia and A. Sharma, "GIS-based road accident black spot analysis in India," International Journal of Geographic Information Science, vol. 34, no. 5, pp. 765–779, 2022.

[7] S. Gupta and R. Mehta, "Comparative analysis of K-Means and DBSCAN clustering for accident datasets," International Journal of Computer Applications, vol. 176, no. 3, pp. 15– 22, 2020.

[8] H. Park and S. Lee, "Hierarchical clustering for traffic accident analysis in South Korea," Journal of Advanced Transportation, vol. 2019, Article ID 8456231, pp. 1–11, 2019.

[9] A. Ahmed and M. Khan, "Geospatial visualization of accident hotspots using machine learning and GIS," Arabian Journal for Science and Engineering, vol. 45, pp. 7651–7664, 2020.

[10] S. Rao, D. Patel, and P. Reddy, "Road accident analysis and visualization using heatmaps: A case study of Telangana," International Journal of Transportation Research, vol. 9, no. 2, pp. 112–121, 2021.