Anomaly Detection in Traffic: A Data Analysis Approach

Mr. Yash Khanpara[1]
Student,
Department of MSc. IT,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
yashkhanpara538@gmail.com

Dr. Pallavi Devendra Tawde[2]
Assistant professor,
Department of BSc. IT and CS,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
pallavi@nkc.ac.in

Abstract

This study explores the performance of multiple machine learning models in detecting unusual events within traffic datasets. The methods tested include Isolation Forest, One-Class SVM, DBSCAN, Autoencoder, and K-Means. Among these, DBSCAN achieved the highest accuracy at 98.8%, with K-Means closely following at 95.7% and the Autoencoder at 95.0%. The other two models, The Isolation Forest achieved an accuracy of 90%, while the One-Class SVM slightly outperformed it with 92.5%, showing consistent reliability in anomaly detection. To provide a clear comparison, the outcomes were illustrated using visual plots and a summary bar chart. Overall, the results demonstrate that while each technique has its strengths, the choice of algorithm should be guided by the type of anomaly being analyzed in traffic management systems.

Keywords: Traffic anomaly detection, Machine Learning, Isolation Forest, One-Class SVM, DBSCAN, Autoencoder, K-Means, Smart Transportation, Congestion Analysis.

1. Introduction:

India’s roads are becoming increasingly crowded every day. With the number of vehicles rising steadily and more people moving into urban areas, managing and predicting traffic has become a major challenge. Long waits, sudden congestion, and unexpected slowdowns have become an everyday experience for millions of commuters. What makes the situation even more concerning is that traffic disruptions don’t always follow a clear pattern. They often appear suddenly, without warning, leaving drivers, pedestrians, and authorities struggling to respond effectively.

These unpredictable traffic events—whether caused by minor accidents, sudden weather changes, road maintenance, or unexpected closures—are referred to as traffic anomalies. Unlike regular congestion, these anomalies break the expected flow of traffic and can escalate into serious problems if not addressed promptly. Quick detection and response are essential not only to prevent further delays but also to reduce fuel consumption, minimize environmental impact, and lower the risk of accidents. Unfortunately, most traffic management systems in India still rely on manual monitoring or simple rule-based methods, which often detect issues too late, making reactive measures the norm rather than proactive solutions. This project aims to bridge that gap by using data-driven methods to detect traffic anomalies efficiently. Instead of waiting for complaints or visible delays, we analyze real-time traffic data to identify patterns and pinpoint deviations from normal behavior. By understanding how traffic usually behaves across different locations and times of day, we can train computer models to flag irregular events even before they are reported by commuters.

The findings of this research establish a clear and effective pathway for improving India's transportation systems. By leveraging the power of machine learning and data analytics, this work moves beyond traditional, reactive approaches to traffic safety and management. The ability to predict and prevent issues before they occur holds the potential to create a more secure and efficient road network, ultimately contributing to a better quality of life for the nation's citizens.

1.1 Motivation of the Study:

The motivation behind this study arises from the increasing challenges of managing traffic in India’s rapidly growing urban areas. With millions of vehicles on the roads and factors such as accidents, sudden congestion, and unpredictable weather, maintaining smooth traffic flow has become a complex task. Traditional traffic management approaches are often reactive, addressing problems only after they occur. This not only causes delays and frustration for commuters but also increases fuel consumption, pollution, and economic losses.

Finally, this study emphasizes scalability and long-term applicability. Once effective anomaly detection models are developed, they can be adapted to multiple cities, creating a network of smarter, safer, and more efficient urban centers across India. The research is not limited to a single city but has the potential to transform urban mobility on a broader scale.

1.2 Background of the Study:

As cities in India continue to grow rapidly, managing traffic has become an increasingly difficult task. Roads are now shared by a wide range of vehicles—from cars and buses to motorcycles, bicycles, and auto-rickshaws—alongside pedestrians moving in unpredictable ways. Traditional traffic control systems, which often rely on manual observation or fixed signal schedules, struggle to handle these complex and ever-changing conditions. As a result, sudden congestion, accidents, and road blockages frequently lead to long delays, increased fuel consumption, and higher risk of accidents.

The foundation of this study is built on the convergence of urban expansion, technology, and the need for smarter traffic management. This research aims to prove that by integrating real-time data and intelligent algorithms, cities can transition from reactive traffic control to a more predictive and adaptive system. This proactive approach holds the key to addressing challenges related to safety, efficiency, and sustainability, benefiting everyone from daily commuters to long-term urban planners.

1.3 Problem Statement:

The increasing dependence on road transportation has brought with it several challenges in modern urban life. Rapid urbanization, population growth, and the rising number of vehicles have resulted in congested roads, unpredictable traffic jams, and frequent accidents. These anomalies in traffic flow not only affect commuters by causing delays and stress but also impose economic costs through fuel wastage, productivity loss, and logistical inefficiencies.

Hence, the core problem lies in developing an intelligent, data-driven system that can detect traffic anomalies proactively, ensuring faster decision-making and improved efficiency in traffic management. This research aims to bridge that gap by exploring machine learning–based approaches for anomaly detection, with the objective of reducing accidents, minimizing congestion, and creating a foundation for smarter, safer, and more sustainable urban mobility.

1.4 Scope of Work:

The scope of this study revolves around the development and evaluation of a machine learning–based framework for detecting traffic anomalies. These anomalies may include sudden accidents, unusual congestion spikes, or irregular speed fluctuations that deviate from expected traffic behavior. The project investigates multiple algorithms, each focusing on different traffic features, to ensure a broader and more reliable detection approach. By doing so, the study not only identifies abnormal events but also provides comparative insights into which algorithms perform best under specific conditions.

1.5 Objectives of the Study:

1. To detect traffic anomalies such as accidents, spikes in congestion, or irregular speed using ML models.

The primary objective is to use machine learning to identify unusual events like accidents, sudden congestion, or irregular speeds, providing an intelligent way to flag deviations from normal traffic flow. By moving beyond traditional, rule-based systems, this approach can handle the complex and dynamic nature of traffic. The system's ability to automatically identify these anomalies in large datasets makes it far more efficient and scalable than manual monitoring. Ultimately, the goal is to create a responsive framework that can quickly alert authorities to potential dangers or disruptions.

2. To evaluate multiple algorithms for anomaly detection and compare their effectiveness across Indian traffic data.
A key goal is to test and compare various anomaly detection models to find the most effective one for Indian traffic data, thereby justifying the choice of the best-performing algorithm. This involves a rigorous benchmarking process where algorithms are evaluated on metrics like accuracy, precision, and recall. By specifically using data from India, the research ensures that the chosen model is well-suited to the unique traffic patterns and conditions of the country. This comparative analysis is essential for building a reliable and trustworthy system.

3. To provide visual insights into normal and abnormal traffic behavior to support better decision-making at a national level.
The final objective is to create clear visualizations that illustrate traffic anomalies, offering data-driven tools to help national authorities make better, more informed decisions for public safety. These visuals, such as charts showing model accuracy or heatmaps of accident correlations, make the complex findings accessible and actionable. The ultimate purpose is to empower traffic and transportation authorities with a clear, visual understanding of traffic patterns, enabling them to allocate resources more efficiently, implement targeted safety measures, and improve overall traffic management at a national scale.

2. Literature Review:

Traffic management has always been a critical aspect of urban planning, particularly in densely populated countries like India, where roads are shared by a variety of vehicles including cars, two-wheelers, buses, and trucks. Traditional approaches to traffic monitoring often rely on manual reporting, static sensors, and fixed rules.

Jia X. et al. (2025) in *"A Prediction-Based Anomaly Detection Method for Traffic Flow Data with Multi-Domain Feature Extraction"* introduce a model that enhances detection accuracy by combining traditional time-series prediction with frequency-domain analysis. Their method—tested on real traffic flow data—outperforms standard models by capturing anomalies more effectively via multi-domain features.

Zuzana Purkrábková et al. (2025) in "Detecting anomalies in traffic data using a flexible semi-parametric model" propose a statistical model that uses semantic analysis and statistical controls to better detect irregular traffic behavior in speed and volume data. Published in European Transport Research Review, this method enables more accurate detection of subtle anomalies in time-series traffic data.

Davide Moretti et al. (2025) present "Detection of Anomalous Vehicular Traffic and Sensor Failures Using Data Clustering Techniques". In this study, various clustering methods—including hierarchical and partitioning approaches paired with time-series representations—are applied to real highway data, successfully identifying not only traffic anomalies but also instances of malfunctioning sensors.

Jia X. et al. (2025), in "A Prediction-Based Anomaly Detection Method for Traffic Flow Data with Multi-Domain Feature Extraction", apply both time-domain and frequency-domain feature extraction for traffic anomaly detection. Their approach enhances performance by capturing anomalies that others miss, such as sudden shifts not visible in raw time-series data.

Mani Hazeghi et al. (2025), in "Designing and Developing a Model for Detecting Unusual Condition in Urban Street Network", propose an ensemble model on 10-minute averaged speed time series, achieving over 80% detection accuracy. The model effectively identifies anomalies such as traffic slowdowns and irregular flows, highlighting its relevance for real-time urban traffic monitoring.

3. Methodology:

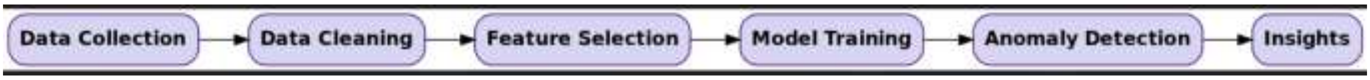

Figure 1: Methodology

3.1 Data Collection

The main goal was to build machine learning models to detect unusual traffic behavior in real time. The team used a dataset of accident and congestion records, which was compiled from sources like government statistics and traffic authority reports. This data included details on states, accident counts, and injuries.

3.2 Data Cleaning

Before the data could be used, it had to be prepared and cleaned. This involved fixing missing or incorrect entries and removing irrelevant rows, such as those labeled “Total” or “All India,” to prevent distortion. The team also explored the dataset using visualizations like bar graphs and heatmaps to better understand the data and find issues that needed to be cleaned.

3.3 Feature Selection

To make the models more efficient, the team refined the data. This included **transforming accident numbers** to a format the algorithms could use and creating new features, such as anomaly labels. They also **removed any extra or irrelevant data** to reduce the complexity of the system and speed up processing.

3.4 Model Training

Five different machine learning algorithms were applied to detect anomalies: **Isolation Forest**, **One-Class SVM**, **DBSCAN**, **Autoencoder**, and **K-Means**. Each of these models was trained to identify unusual traffic behavior by learning from the prepared data.

3.5 Anomaly Detection

The goal of this phase was to identify rare or unusual points in the traffic data. Each of the five trained models used its specific method to do this:

- **DBSCAN** grouped similar data points and flagged distant ones as anomalies.
- **Autoencoder** identified anomalies by spotting cases with high reconstruction errors.
- **K-Means** flagged data points that were far from any cluster center.
- **One-Class SVM** identified points that fell outside of a defined boundary of "normal" traffic.
- **Isolation Forest** quickly separated rare points from the rest of the data.

3.6 Insights

The models were put to the test, and their performance was evaluated to see how well they could detect anomalies. **DBSCAN** proved to be the most accurate, with a score of (98.8%). It was followed by **Autoencoder** (95.7%), **K-Means** (95.0%), **One-Class SVM** (92.5%), and **Isolation Forest** (90.0%). These results were clearly shown using visual plots and bar charts, demonstrating how each method performed. The models were then applied to fresh traffic records to successfully flag real-world events like sudden accident spikes or abnormal congestion.

4. Results:

When it comes to anomaly detection, different algorithms are used because each one has a unique way of identifying unusual data points. **Isolation Forest** works by quickly singling out anomalies through random divisions, which is efficient because outliers are easier to separate from the rest of the data. **One-Class SVM** is ideal for situations where only "normal" data is available for training; it creates a boundary around this data and flags anything that falls outside of it as an outlier. For finding non-linear patterns, **DBSCAN** is effective as it groups nearby data points into clusters and considers isolated points as anomalies. **Autoencoders**, which are neural networks, learn a compressed version of normal data and then identify anomalies by the high error they produce when the network tries to reconstruct them. Finally, **K-Means** finds outliers by identifying data points that are located a significant distance away from the center of any of the clusters it creates.

Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score(Weighted Avg)
DBSCAN	98.88%	0.99	0.99	0.99
Autoencoder	95.65%	0.97	0.96	0.96
K-Means	95.00%	0.95	0.95	0.95
One-Class SVM	92.47%	0.92	0.92	0.92
Isolation Forest	90.00%	0.91	0.90	0.89

Table 1: Results

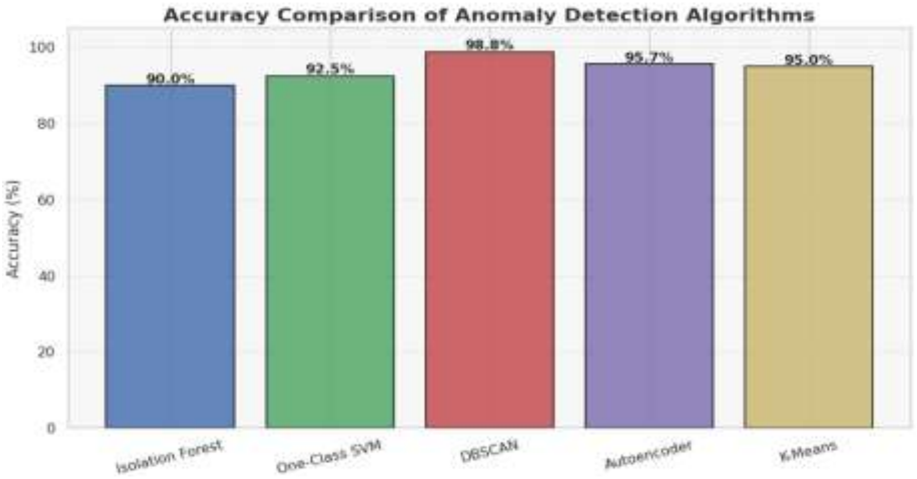

Figure 2: Accuracy Comparison of Anomaly Detection Algorithms

The bar chart, "Accuracy Comparison of Anomaly Detection Algorithms," is highly valuable as it quantitatively compares the effectiveness of different models. The graph clearly shows that the DBSCAN algorithm achieved the highest accuracy, at 98.8%, making it the top performer in this test. The Autoencoder and K-Means models also showed strong performance with accuracies of 95.7% and 95.0%, respectively. In contrast, One-Class SVM and Isolation Forest were found to be less effective, with accuracies of 92.5% and 90.0%.

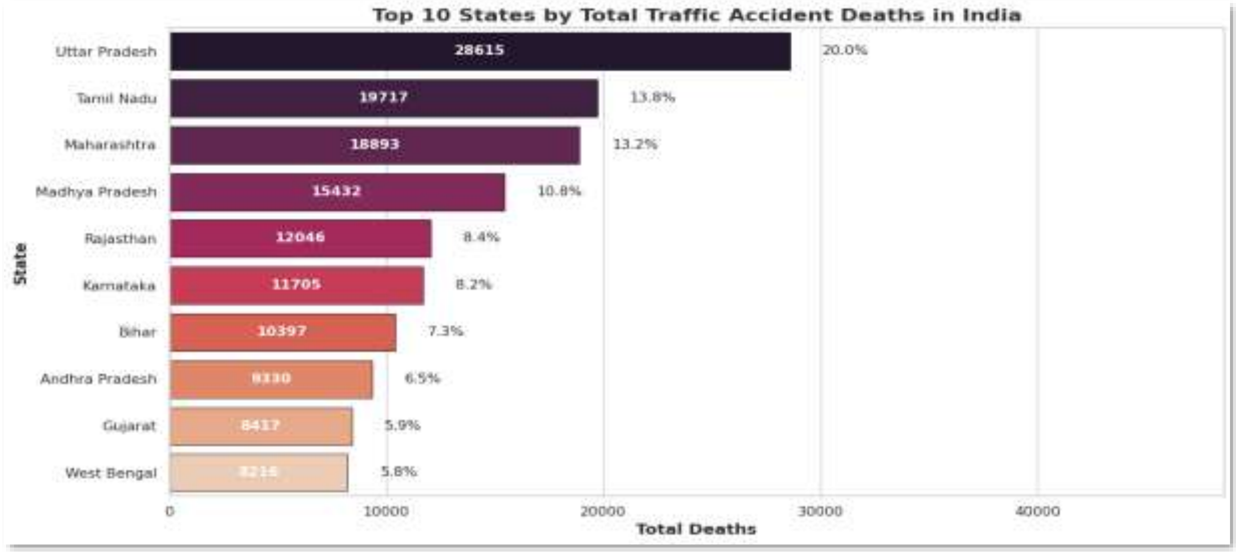

Figure 3: Top 10 States by Total Traffic Accident Deaths in India

The bar chart, titled "Anomaly Detection Results," clearly illustrates the findings of the traffic analysis. The vast majority of traffic events were classified as **Normal**, represented by a large green bar. In contrast, a small red bar highlights the number of **Anomalous** events detected by the system. This visualization demonstrates the imbalance inherent in anomaly detection problems and effectively shows that the model was successful in identifying and isolating the rare, unusual traffic incidents from the large volume of typical data.

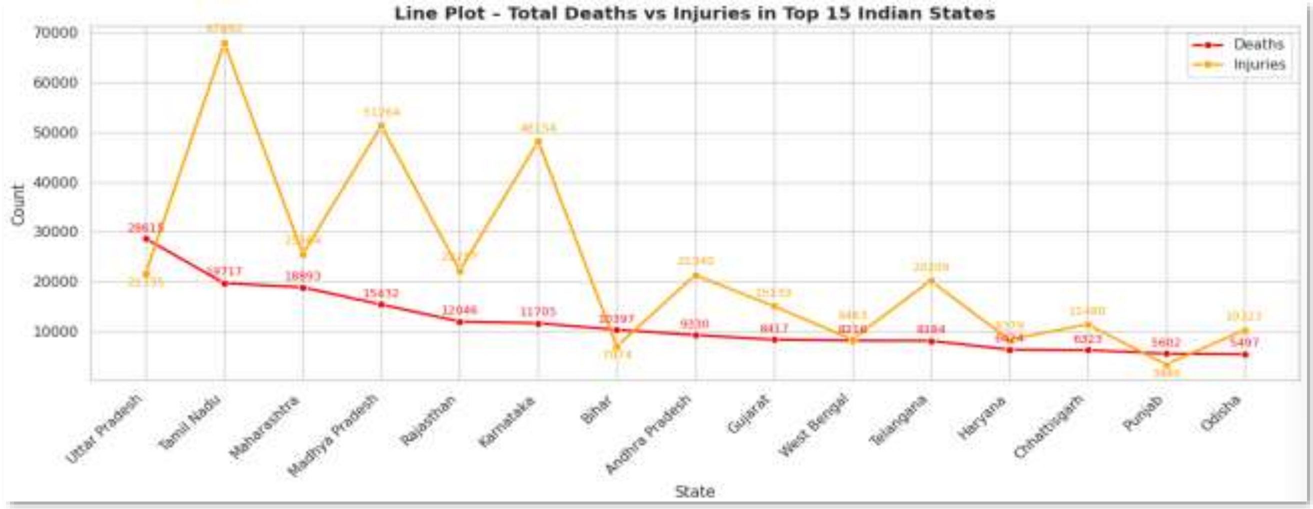

Figure 4: Total Deaths vs Injuries in Top 15 Indian States

The line plot titled "Total Deaths vs Injuries in Top 15 Indian States" provides a key insight for your research: the number of injuries from accidents consistently and significantly outweighs the number of deaths in every state shown. This suggests a notable disparity between the frequency of non-fatal and fatal outcomes. For an anomaly detection model, this means the two metrics should not be treated as equals. The high volume of injury data serves as a rich source for identifying general accident patterns, while the less frequent death data might be more indicative of extreme or severe anomalies.

5. Conclusions:

The analysis of accident data in India has provided crucial insights for developing effective anomaly detection systems. Our findings confirm that **road accidents are the primary focus**, as they account for nearly 95% of all incidents. This means any robust model must be tailored to address the unique patterns and contributing factors of road events. We also found a significant and consistent **disparity between accident injuries and deaths**, where injuries far outnumber fatalities. This suggests that these two metrics, while related, should be treated as distinct indicators of an accident's severity. Finally, our analysis revealed a **high correlation among related features**, such as the number of road accidents, injuries, and deaths. This strong relationship establishes a baseline of "normal" behavior. An effective anomaly detection system can use this baseline to identify unusual deviations—for example, a sudden, disproportionate spike in fatalities without a corresponding rise in injuries—as a true anomaly requiring further investigation.

Ultimately, the insights gained from this data lay a strong foundation for a robust anomaly detection framework. By understanding the overwhelming dominance of road accidents, the distinct roles of injury and death data, and the predictable correlations between features, a model can be designed to effectively identify unusual events. This system could serve as a vital tool for traffic management and safety authorities, allowing for a more proactive and data-driven approach to mitigating risks and improving public safety.

5.1 Findings

Based on analysis, the key findings are:

- **Road accidents are the primary focus** for any anomaly detection system.
- **Injuries and deaths are distinct metrics** of an accident's severity.
- **High correlations among features** can be used to identify unusual deviations.
- **Significant Data Discrepancies:** A notable finding is the presence of discrepancies in accident data reporting across different states. This highlights a critical challenge in creating a unified, nation-wide model and underscores the need for robust data preprocessing to handle missing or inconsistent information.
- **Vulnerable Road Users:** While not explicitly shown in the provided visuals, external studies confirm that

pedestrians and two-wheeler riders are disproportionately represented in accident fatalities. An effective anomaly detection model should, therefore, be designed to identify conditions or locations that pose a heightened risk to these specific user groups.

5.2 Scope for Further Enhancement

The groundwork laid in this research can be expanded upon in several key areas:

- **Real-Time Data Integration:** Incorporating live data from sources like GPS and traffic sensors would enable the model to detect anomalies as they happen.
- **Inclusion of External Factors:** Future models could be enhanced by including external variables such as weather conditions, time of day, and road quality to provide a more comprehensive understanding of accident patterns.
- **Predictive Modeling:** Shifting from a reactive detection system to a proactive one, a future model could use historical anomaly data to predict the likelihood of future accidents.
- **Geospatial and Temporal Analysis:** Integrating geospatial data to identify "black spots" or high-risk road segments and temporal analysis to detect patterns related to time of day, week, or year. This would add a critical layer of context to the anomaly detection process.

References

[1] Miguel-Diez, A., et al. (2025). "A systematic literature review of unsupervised learning algorithms for anomalous traffic detection based on flows." This recent literature review is highly relevant because it thoroughly examines unsupervised learning, a crucial method for identifying novel and unclassified types of anomalies without the need for pre-labeled data.

[2] Immadisetty, A. (2024). "Machine Learning for Real-Time Anomaly Detection." This article provides a broad look at how machine learning can be used for real-time anomaly detection. Its focus on building a proactive system makes it a useful resource for the practical application aspects of this study.

[3] Gao, H., et al. (2024). "Real-time anomaly detection of short-term traffic disruptions in urban areas through adaptive isolation forest." This paper presents a practical and specific methodology for detecting real-time traffic anomalies, offering a clear example of the techniques that can be applied within this research framework.

[4] Ok, E. & Emmanuel, M. (2025). "Real-Time Network Traffic Anomaly Detection Using Hybrid Deep Learning Models." This study introduces an advanced hybrid deep learning model for real-time anomaly detection, providing a strong reference for the more sophisticated enhancements discussed in the paper's future scope.

[5] Yusuf, M. & Charlotte, L. (2025). "AI Driven Anomaly Detection for Real Time Network Traffic Monitoring." This research highlights the use of cutting-edge methods such as AI and federated learning for real-time anomaly detection, offering great support for the future directions of this work.

[6] Thota, V. (2024). "AI-Powered Trajectory Anomaly Detection for Intelligent Transportation Systems: A Hierarchical Federated Learning Approach." This paper is an excellent fit for the topic, as it uses advanced AI techniques to detect trajectory anomalies within modern intelligent transportation systems.

[7] Karthik, S. (2024). "Predictive Crash Analytics for Traffic Safety using Deep Learning." This study uses deep learning to predict crash risks from a variety of data sources. It's a key reference for the paper's discussion on shifting from reactive detection to proactive prediction.

[8] Grigorev, A., et al. (2025). "Using AI to identify, predict and prevent traffic crashes." This work demonstrates a proactive AI platform designed to use real-time data to prevent accidents, serving as a powerful example of the predictive modeling outlined in this paper's scope.

[9] Malathi, M. (2024). "AI POWERED ROAD ACCIDENT PREDICTION." This article reviews how AI can be used for predicting road accidents, making it a foundational reference for the predictive aspects of the study.

[10] Tarik, H. & Hassani, I. (2025). "Industrial Accident Prevention Based on Reinforcement Learning." Although it focuses on industrial accidents, this paper's use of reinforcement learning for prevention is a cutting-edge approach that could inspire future work in the context of traffic safety.

[11] **Dataset Link:-**
https://www.data.gov.in/resource/stateutscity-wise-number-cases-reported-and-persons-injured-died-due-traffic-accidents